Федеральное агентство по образованию
ГОУ ВПО «Российский государственный
профессионально-педагогический университет»
Уральское отделение Российской академии образования
Академия профессионального образования

В. А. Густомесов

ЭКОНОМЕТРИКА

Учебное пособие

Рекомендовано УМО по математике
педвузов Волго-Вятского региона
в качестве учебного пособия
для студентов экономических специальностей
высших учебных заведений

Екатеринбург
2007

УДК 330.4 (075.8)
ББК У.в 631я73-1
Г 96

Густомесов В.А. Эконометрика [Текст]: учеб. пособие/ В.А. Густомесов. Екатеринбург: Изд-во ГОУ ВПО «Рос. гос. проф.-пед. ун-т», 2007. 128 с.
ISBN 978-5-8050-0216-9

В учебном пособии по дисциплине «Эконометрика» рассмотрены основные методы эконометрического моделирования, основанные на регрессионном анализе: линейная регрессия, нелинейная регрессия, временные ряды, системы одновременных уравнений. Теоретический материал иллюстрируется примерами построения и исследования эконометрических моделей.

Издание предназначено для студентов специальностей 060100 Экономическая теория и 351400 Прикладная информатика (в экономике).

Рецензенты: доктор физико-математических наук, профессор А.Ф. Клейменов (Институт математики и механики УрО РАН); кандидат физико-математических наук, доцент Л.С. Чебыкин (ГОУ ВПО «Российский государственный профессионально-педагогический университет»)

ISBN 978-5-8050-0216-9

Оглавление

Предисловие

Учебное пособие «Эконометрика» предназначено для студентов специальностей 060100 Экономическая теория и 351400 Прикладная информатика (в экономике). Оно соответствует государственным образовательным стандартам для этих специальностей.

Эконометрика в настоящее время является одной из базовых учебных дисциплин экономического образования. Экономисту зачастую приходится изучать реальные экономические объекты и процессы на основе статистических данных. Такое исследование и составляет содержание эконометрики.

Для понимания курса «Эконометрика» требуется владение базовыми знаниями по основным разделам математики, особенно по теории вероятностей и математической статистике [1; 5; 21], линейной алгебре, дифференциальному исчислению.

Настоящее пособие состоит из введения, четырех глав и приложения.

Во введении раскрываются предмет эконометрики и особенности эконометрического моделирования.

В первой главе подробно описаны линейные регрессионные модели, составляющие ядро эконометрики.

Вторая глава посвящена нелинейным моделям регрессии. Отмечены способы преобразования нелинейных моделей в линейные.

В третьей главе изучены временные ряды – специфические регрессионные модели, предназначенные для описания экономических процессов.

В четвертой главе рассмотрены особенности исследования сложных эконометрических моделей, описываемых системами взаимосвязанных (одновременных) уравнений.

Приложение содержит сведения из теории матриц, используемые в основном тексте пособия.

Теоретический материал, изложенный в пособии, иллюстрируется

примерами построения и исследования конкретных эконометрических моделей.

При написании пособия в определенной степени учтен опыт преподавания эконометрики на кафедре высшей математики ГОУ ВПО «Российский государственный профессионально-педагогический университет», где сложилась традиция широкого применения в эконометрическом анализе аппарата выборочных ковариаций [15; 18].

В связи с небольшим объемом часов, отводимых в учебных планах на изучение эконометрики, материал пособия сконцентрирован вокруг основной идеи, суть которой заключается в исследовании эконометрических моделей средствами регрессионного анализа. За рамками пособия остался, например, подход к анализу временных рядов, связанный с теорией случайных функций.

Автор признателен рецензентам – профессору А.Ф. Клейменову и доценту Л.С. Чебыкину – за полезные советы.

Введение

Термин «эконометрика» (иногда говорят «эконометрия») буквально означает «измерения в экономике».

Эконометрика описывает количественные закономерности и взаимосвязи конкретных экономических объектов посредством эконометрических моделей. Она опирается на экономическую теорию и различные разделы математики, прежде всего на теорию вероятностей и математическую статистику.

Рыночная экономика требует эффективного использования имеющейся информации о результатах хозяйственной деятельности. Для принятия управленческих решений менеджеру необходимо знать систему связей между экономическими факторами, выделять из них основные связи, выдвигать обоснованные экономические прогнозы и гипотезы. Такие задачи решаются посредством построения и исследования эконометрических моделей.

Эконометрические модели являются специфическим классом вероятностно-статистических *математических моделей.* Введем понятие математической модели и укажем особенности эконометрического моделирования.

Математика изучает абстрактные математические объекты, напрямую не связанные с материальным миром. В то же время она является важным средством познания явлений внешнего мира (реальных объектов), изучая их математические модели.

Математическая модель – *приближенное описание реального объекта, выраженное с помощью математической символики.*

Процесс математического моделирования в большинстве случаев можно разбить на следующие основные шаги (рисунок):

А. *Построение* математической модели изучаемого объекта на основе эмпирических фактов, законов соответствующей области науки и гипотез.

Б. *Исследование* построенной модели средствами математики.

В. *Интерпретация* результатов исследования модели в терминах объекта.

Г. *Сравнение* свойств модели и объекта, при необходимости – уточнение математической модели или построение новой математической модели.

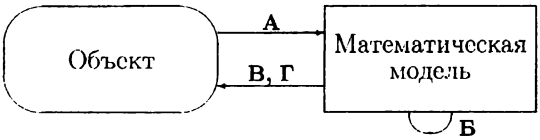


Схема математического моделирования

При математическом моделировании реального объекта обычно выделяют его основные, наиболее существенные факторы (переменные) и изучают связи лишь между ними.

Математические модели можно разделить на *детерминированные* и *вероятностно-статистические.*

Детерминированные модели предполагают жесткие, функциональные связи между переменными. Назовем важные классы детерминированных моделей:

- модели, использующие аппарат дифференциальных уравнений, но не содержащие случайных составляющих;
- модели линейного программирования.

Вероятностно-статистические (стохастические) модели обязательно включают *случайные* воздействия на переменные; при их исследовании используется аппарат теории вероятностей и математической статистики.

Эконометрические модели по своей природе являются стохастическими и строятся на основе статистических данных (выборки) об исследуемом экономическом объекте. Часто бывает доступна и другая априорная информация об объекте, представленная в форме закономерностей экономической теории, ограничений, гипотез о характере случайных воздействий.

Эконометрическая модель обычно записывается в виде системы ал-

гебраических уравнений, содержащих выбранные переменные и неизвестные параметры, которые оценивают по статистическим данным. Модель может состоять и из одного уравнения.

Процесс эконометрического моделирования во многих случаях разбивают на следующие этапы:

1) *Постановочный этап.* Он включает определение *целей* моделирования; выделение переменных модели; разбиение выделенных переменных на *объясняющие*, независимые переменные (регрессоры) и *объясняемые*, зависимые переменные.

2) *Априорный (предмодельный) этап.* На этом этапе осуществляется анализ априорной информации об изучаемом экономическом объекте.

3) *Информационно-статистический этап.* Он посвящен сбору статистической информации об объекте, т.е. регистрации значений выделенных переменных объекта.

4) *Этап спецификации модели.* Данный этап состоит в *выборе структуры* эконометрической модели, т. е. *общего вида* соотношений, связывающих выделенные переменные.

5) *Этап идентификации модели.* На этом этапе производится статистическое оценивание ее параметров, выявляются статистические свойства найденных оценок.

6) *Этап верификации (проверки истинности) модели.* Здесь осуществляются сопоставление свойств модели и исходной информации об объекте, выявление соответствия модели наблюдениям, оценка точности модели. Если в результате установлены недопустимые расхождения между объектом и моделью, то необходима корректировка модели; приходится возвращаться к 4-му, а иногда и к 1-му этапу.

Мы в основном будем обсуждать проблемы, связанные с реализацией 4–6-го этапов.

Этапы 1–4 соответствуют шагу **А** общей схемы математического моделирования и конкретизируют процедуру построения эконометрической модели. Этап 5 отвечает шагам **Б, В**; здесь одновременно производятся и исследование модели, и интерпретация результатов моделирования. Последний, 6-й, этап соответствует шагу **Г**.

Основными целями эконометрического моделирования являются:

- *анализ* исследуемого экономического объекта;
- *прогноз* значений его зависимых переменных при фиксированных

значениях независимых факторов;

- *имитация* различных возможных сценариев развития объекта;
- *выработка управленческих решений.*

Эконометрические модели строятся для описания экономических объектов разных *уровней иерархии* и различного *профиля.*

В качестве основных уровней иерархии принимаем *макроуровень* (мировое сообщество, страна в целом), *мезоуровень*[1] (регионы, отрасли, корпорации) и *микроуровень* (семья, предприятие, фирма).

Основными профилями эконометрических моделей выступают проблемы рынка, инвестиционной, финансовой, социальной политики, ценообразования, распределительных отношений.

В эконометрике используются два основных типа статистических данных: 1) *пространственные* данные (пространственная выборка), 2) *временны́е* данные (временна́я выборка).

Пространственные данные характеризуют *различные* однотипные объекты в определенный момент времени. В этом случае изучаемый экономический объект – объединение этих однотипных объектов.

Временные данные характеризуют экономический объект в *различные* моменты времени (например, данные об ежедневном курсе валют).

Выделим следующие основные типы эконометрических моделей:

1) *регрессионные модели* с одним уравнением, описывающие зависимость одной объясняемой переменной от одного или нескольких регрессоров; они обязательно содержат случайные составляющие;

2) *модели динамических (временны́х) рядов* – специфические регрессионные модели, связанные с временными данными;

3) *системы одновременных уравнений*, включающие регрессионные уравнения и тождества.

Основное внимание мы уделим различным регрессионным моделям. Исследование таких моделей проводится с позиций математической статистики. Для оценки параметров моделей часто применяется метод наименьших квадратов (МНК).

Эконометрика как самостоятельная наука возникла в 30-е г. XX в. на стыке экономической теории, статистических и математических методов. К этому времени значительное развитие получили и экономиче-

[1]От греч. «mesos» – средний.

ская теория, и математическая статистика. Метод наименьших квадратов был разработан немецким математиком К. Гауссом на рубеже XVIII и XIX вв. и вначале применялся им для обработки результатов астрономических и геодезических наблюдений. Строгое математико-статистическое обоснование метода было дано в трудах русских математиков А.А.Маркова (1898) и А.Н. Колмогорова (1946).

Считается, что знаковым шагом, знаменующим появление эконометрики, явилось создание в 1930 г. Эконометрического общества, куда вошли американские и европейские ученые. Важную роль в создании молодой науки сыграл журнал «Econometrica», издающийся с 1933 г. (его первый редактор – Рагнар Фриш).

В эконометрической практике широко используются традиционные математико-статистические методы. Возникли и новые, собственно эконометрические направления, в том числе теория систем одновременных уравнений, разработанная Трюгге Хаавельмо.

Теперь методы эконометрики относятся к основным методам автоматизации экономических расчетов. Свидетельством признания эконометрики является то обстоятельство, что многие Нобелевские премии по экономике присуждаются за достижения именно в области эконометрики. В числе лауреатов премии – Р. Фриш, Т. Хаавельмо, автор известной макроэкономической модели Л. Клейн.

1. Линейные регрессионные модели

1.1. Регрессионные модели

Регрессионные модели являются основными моделями эконометрики.

Укажем *общий вид* регрессионной модели. Будем считать, что 1-3-й этапы эконометрического моделирования уже проведены. В частности, выделены *существенные переменные* изучаемого экономического объекта

$$x_1, \ldots, x_m, y \quad (m \in \mathbb{N}).$$

Они состоят из m *независимых* переменных $x_1, \ldots, x_m$ (их называют также *регрессорами, объясняющими переменными*) и из *одной зависимой (объясняемой)* переменной y. Впрочем, при исследовании регрессионных моделей может быть проведена коррекция набора регрессоров (см. подп. 1.3.4).

При построении модели учитывается выборка статистических данных об объекте, состоящая из n наблюдений над этим объектом:

$$\begin{gathered}(x_{11}, x_{12}, \ldots, x_{1m}, y_1),\ (x_{21}, x_{22}, \ldots, x_{2m}, y_2),\ \ldots, \\ (x_{i1}, x_{i2}, \ldots, x_{im}, y_i),\ \ldots,\ (x_{n1}, x_{n2}, \ldots, x_{nm}, y_n).\end{gathered} \tag{1.1}$$

Здесь y_i – значение объясняемой переменной y, $x_{i1}, x_{i2}, \ldots, x_{in}$ – значения регрессоров $x_1, x_2, \ldots, x_m$ при i-м наблюдении ($i \in \overline{1,n}$). Число наблюдений n весьма велико и существенно больше m числа объясняющих переменных.

Выборка (1.1) задает в $(m+1)$-мерном пространстве $\mathbb{R}^{m+1}$ переменных $x_1, x_2, \ldots, x_m, y$ множество, состоящее из n точек. Это множество называют *корреляционным полем* (полем рассеивания). Однако оно будет достаточно наглядным лишь в простейшем случае одного регрессора ($m = 1$); тогда корреляционное поле есть множество точек плоскости.

Как известно из курса теории вероятностей и математической статистики [1, с. 622; 5, с. 173], регрессия y на $x_1, \ldots, x_m$ определяется как условное среднее значение (условное математическое ожидание) случайной величины y, если переменные x_l $(1 \leqslant l \leqslant m)$ принимают фиксированные значения x_l^*:

$$M(y|\, x_1 = x_1^*, \ldots, x_m = x_m^*).$$

Это неслучайная функция m переменных $x_1^*, \ldots, x_m^*$:

$$M(y|\, x_1 = x_1^*, \ldots, x_m = x_m^*) = f(x_1^*, \ldots, x_m^*).$$

В дальнейшем мы аргументы функции f будем обозначать через $x_1, \ldots, x_m$. Следовательно, $f(x_1, \ldots, x_m)$ – условное математическое ожидание зависимой переменной y в предположении, что регрессоры принимают значения $x_1, \ldots, x_m$.

Уравнение регрессии запишем в виде

$$y = f(x_1, \ldots, x_m) + \varepsilon. \tag{1.2}$$

В выбранной модели (1.2) случайная величина y представима в виде *суммы функции регрессии*

$$f(x_1, \ldots, x_m) \tag{1.3}$$

и *случайного слагаемого* ε. Слагаемое ε отражает влияние факторов, не включенных в модель, а также ошибок измерения. Его называют *возмущением* или *ошибкой регрессии*.

Функция регрессии (1.3) есть функция некоторого фиксированного класса. Выбор этого класса задает *структуру* регрессионной модели и проводится на 4-м этапе (спецификации) эконометрического моделирования. При определении структуры модели учитываются характер корреляционного поля, априорная информация об объекте и, возможно, опыт предыдущих эконометрических исследований. Целесообразно выбирать простые регрессионные модели, желательно *линейные*. Как образно сказал А. Эйнштейн о математическом моделировании, модели должны быть «настолько простыми, насколько возможно, но не проще» (цит. по: [14, с. 26]).

Функция регрессии (1.3) обычно зависит от некоторых *параметров (коэффициентов)*, которые неизвестны. *Точные* их значения, как правило, и не могут быть найдены. Исследование регрессионной модели (1.2)

проводится в рамках подхода, характерного для математической статистики. По выборке (1.1) находятся *статистические оценки* неизвестных параметров, тем самым определяется статистическая оценка

$$\widehat{y} = \widehat{f}(x_1, \ldots, x_m)$$

функции регрессии (1.3). Оценивание параметров обычно проводится методом наименьших квадратов, который позволяет в определенном смысле наилучшим образом распорядиться наблюдениями (1.1). После нахождения оценок изучаются их статистические свойства, проводится сопоставление свойств объекта и построенной модели.

Особо выделим простой случай *одного* регрессора ($m = 1$). Регрессор удобно обозначить символом x, опуская нижний индекс. Тогда регрессионная модель (1.2) запишется следующим образом:

$$y = f(x) + \varepsilon. \tag{1.4}$$

Модель (1.4) называется *моделью парной регрессии*. Она связывает пару переменных x, y.

Специфика парной модели состоит в том, что при выборе структуры модели, т.е. функции $f(x)$, можно эффективно учитывать характер корреляционного поля, являющегося множеством точек плоскости (x, y). Действительно, в случае $m = 1$ выборка (1.1) представляет собой множество пар

$$(x_1, y_1), \ldots, (x_n, y_n). \tag{1.5}$$

В случае нескольких регрессоров ($m > 1$) модель (1.2) называют моделью *множественной* регрессии. Понятно, что она является обобщением модели (1.4).

Чтобы регрессионная модель содержала статистическую информацию об объекте моделирования, подставим в формулу (1.2) данные выборки:

$$y_i = f(x_{i1}, \ldots, x_{im}) + \varepsilon_i, \quad i \in \overline{1, n}. \tag{1.6}$$

Здесь случайные величины ε_i – *ошибки регрессии*. Из формулы (1.6) следует, что наблюдаемые значения y_i зависимой переменной y также случайные величины. Однако, в отличие от y_i, ошибки регрессии ε_i в принципе не могут быть наблюдаемы.

Модель (1.6) называют *регрессионной моделью в наблюдениях*. Она является основной рабочей моделью.

Переходим к систематическому изложению теории *линейных регрессионных моделей*, отвечающих *линейным* функциям регрессии

$$f(x_1, \ldots, x_m) = b_0 + b_1 x_1 + \ldots + b_m x_m. \tag{1.7}$$

Здесь $b_0, b_1, \ldots, b_m$ – неизвестные коэффициенты линейной регрессии, которые необходимо статистически оценить.

1.2. Точечные оценки параметров линейной регрессии методом наименьших квадратов

Исследование модели линейной регрессии начнем с ключевого этапа – нахождения точечных оценок коэффициентов регрессии посредством метода наименьших квадратов. Изучим отдельно случаи парной регрессии (подп. 1.2.1) и множественной регрессии (подп. 1.2.3). Обозримое решение задачи оценивания в общем случае дается с использованием матричного аппарата, поэтому подп. 1.2.2 посвящен преобразованию линейной регрессионной модели в наблюдениях к матричной форме.

1.2.1. Парная линейная регрессия

Вначале рассмотрим простейшую из линейных регрессионных моделей – *парную* линейную регрессионную модель с *одной* объясняющей переменной x. Тогда

$$f(x) = b_0 + b_1 x \tag{1.8}$$

есть функция регрессии; исходная модель (1.2) запишется в виде

$$y = b_0 + b_1 x + \varepsilon, \tag{1.9}$$

а модель в наблюдениях (1.6) – в виде

$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad (i \in \overline{1,n}). \tag{1.10}$$

Парная линейная модель (1.9) применима, когда моделируемый экономический объект содержит *доминирующий фактор* x, влиянием остальных факторов можно пренебречь, а корреляционное поле (1.5) в основном ориентировано вдоль некоторой прямой.

По выборке (1.5) найдем в определенном смысле наилучшие статистические точечные оценки $\widehat{b}_0, \widehat{b}_1$ неизвестных коэффициентов b_0, b_1.

Будем считать, что в выборке (1.5) имеются *различные* значения как регрессора x, так и зависимой переменной y.

Введем выборочную оценку линейной функции регрессии (1.8) – функцию

$$\widehat{y} = \widehat{f}(x) = \widehat{b}_0 + \widehat{b}_1 x. \tag{1.11}$$

Ее графиком является (выборочная) прямая регрессии

$$L: \quad y = \widehat{b}_0 + \widehat{b}_1 x. \tag{1.12}$$

Прямая регрессии L полностью задается коэффициентами $\widehat{b}_0, \widehat{b}_1$. Наша задача – определить их так, чтобы прямая L была наиболее «близкой» к точкам корреляционного поля. Зададим меру этой «близости».

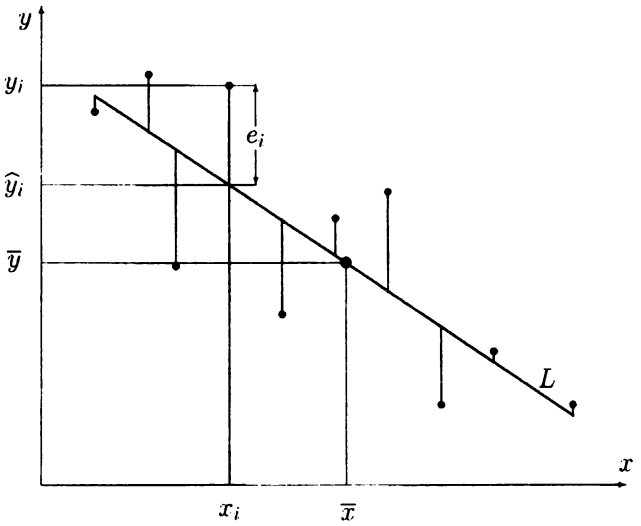


Рис. 1.1. Парная линейная регрессия – метод наименьших квадратов

Рассмотрим разности

$$e_i = y_i - \widehat{y}_i, \tag{1.13}$$

где

$$\widehat{y}_i = \widehat{f}(x_i) = \widehat{b}_0 + \widehat{b}_1 x_i. \tag{1.14}$$

Разности (1.13) - наблюдаемые значения ошибок регрессии ε_i. Их называют *остатками регрессии.*

Каждая разность e_i геометрически задает *отклонение ординаты i-й точки выборки от прямой регрессии* L (вдоль прямой $x = x_i$) (рис. 1.1). Согласно методу наименьших квадратов (МНК) прямую L следует выбрать так, чтобы сумма квадратов всех отклонений

$$\sum_{i=1}^{n} e_i^2 \tag{1.15}$$

была наименьшей. Поэтому МНК означает решение экстремальной задачи

$$\sum_{i=1}^{n} e_i^2 \to \min.$$

Удобно сумму квадратов рассматривать как функцию двух переменных $\widehat{b}_0, \widehat{b}_1$:

$$Q(\widehat{b}_0, \widehat{b}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{b}_0 - \widehat{b}_1 x_i)^2 = \sum_{i=1}^{n} (\widehat{b}_0 + \widehat{b}_1 x_i - y_i)^2.$$

Тем самым приходим к задаче на минимум этой функции

$$Q(\widehat{b}_0, \widehat{b}_1) \to \min. \tag{1.16}$$

Для решения задачи (1.16) убедимся, что существует одна критическая точка функции $Q(\widehat{b}_0, \widehat{b}_1)$, являюшаяся точкой ее минимума.

Критические точки функции являются решениями (относительно коэффициентов $\widehat{b}_0, \widehat{b}_1$) системы уравнений

$$\begin{cases} \dfrac{\partial Q}{\partial \widehat{b}_0} = 0, \\ \dfrac{\partial Q}{\partial \widehat{b}_1} = 0. \end{cases} \tag{1.17}$$

Найдем и преобразуем частные производные первого порядка функции Q :

$$\frac{\partial Q}{\partial \widehat{b}_0} = \sum_{i=1}^{n} \frac{\partial (\widehat{b}_0 + \widehat{b}_1 x_i - y_i)^2}{\partial \widehat{b}_0} =$$

$$= 2\sum_{i=1}^{n}(\widehat{b}_0 + \widehat{b}_1 x_i - y_i) = 2\left(\widehat{b}_0 n + \widehat{b}_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right);$$

$$\frac{\partial Q}{\partial \widehat{b}_1} = 2\sum_{i=1}^{n}(\widehat{b}_0 + \widehat{b}_1 x_i - y_i)x_i = 2\left(\widehat{b}_0 \sum_{i=1}^{n} x_i + \widehat{b}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i\right).$$

Поэтому система (1.17) может быть записана следующим образом:

$$\begin{cases} \widehat{b}_0 n \quad + \widehat{b}_1 \sum\limits_{i=1}^{n} x_i \;\; = \sum\limits_{i=1}^{n} y_i, \\ \widehat{b}_0 \sum\limits_{i=1}^{n} x_i \;\; + \widehat{b}_1 \sum\limits_{i=1}^{n} x_i^2 \;\; = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{1.18}$$

Со времен К. Гаусса систему (1.18) называют системой *нормальных уравнений.*

Введем средние величины, связанные с переменными x, y :

$$\overline{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n},\ \overline{x^2} = \frac{\sum\limits_{i=1}^{n} x_i^2}{n},\ \overline{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n},\ \overline{xy} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{n},\ \overline{y^2} = \frac{\sum\limits_{i=1}^{n} y_i^2}{n}. \tag{1.19}$$

Разделив уравнения системы (1.18) на n, приведем ее к виду

$$\begin{cases} \widehat{b}_0 \quad + \overline{x} \cdot \widehat{b}_1 \;\; = \overline{y}, \\ \overline{x} \cdot \widehat{b}_0 \;\; + \overline{x^2} \cdot \widehat{b}_1 \;\; = \overline{xy}. \end{cases} \tag{1.20}$$

Важно, что уравнения системы (1.20) являются *линейными.* Решим ее *методом подстановки.* Из первого уравнения системы выразим $\widehat{b}_0$ через $\widehat{b}_1$:

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \cdot \overline{x}. \tag{1.21}$$

Теперь преобразуем второе уравнение:

$$\overline{x} \cdot \overline{y} - \widehat{b}_1 (\overline{x})^2 + \widehat{b}_1 \overline{x^2} = \overline{xy},$$

откуда находим

$$\widehat{b}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}. \tag{1.22}$$

Таким образом, функция $Q(\widehat{b}_0, \widehat{b}_1)$ имеет единственную критическую точку, координаты которой определяются формулами

(1.21), (1.22). Выясним, что в этой точке достигается *минимум* функции. Определим ее частные производные второго порядка:

$$\frac{\partial^2 Q}{\partial \widehat{b}_0^2} = 2n > 0, \quad \frac{\partial^2 Q}{\partial \widehat{b}_0 \partial \widehat{b}_1} = 2\sum_{i=1}^{n} x_i, \quad \frac{\partial^2 Q}{\partial \widehat{b}_1^2} = 2\sum_{i=1}^{n} x_i^2.$$

В соответствии с достаточным условием экстремума функции двух переменных находим

$$\Delta(\widehat{b}_0, \widehat{b}_1) = \frac{\partial^2 Q}{\partial \widehat{b}_0^2} \cdot \frac{\partial^2 Q}{\partial \widehat{b}_1^2} - \left(\frac{\partial^2 Q}{\partial \widehat{b}_0 \partial \widehat{b}_1}\right)^2 = 4\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) =$$

$$= 4n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right) \overset{(1.19)}{=} 4n^2(\overline{x^2} - (\overline{x})^2) > 0.$$

В конце преобразований учтено, что

$$\overline{x^2} - (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 > 0.$$

Так как в найденной критической точке значения функций Δ, $\dfrac{\partial^2 Q}{\partial \widehat{b}_0^2}$ положительны, то эта точка является точкой минимума функции $Q(\widehat{b}_0, \widehat{b}_1)$ и единственным решением экстремальной задачи (1.16).

Теперь определяется и прямая регрессии L (1.12). Эта прямая проходит через точку $(\overline{x}, \overline{y})$. В самом деле,

$$\overline{y} \overset{(1.21)}{=} \widehat{b}_0 + \widehat{b}_1\overline{x}.$$

Поэтому уравнение прямой L можно записать, используя лишь коэффициент $\widehat{b}_1$:

$$L: \quad y = \overline{y} + \widehat{b}_1(x - \overline{x}). \tag{1.23}$$

Так как прямая L является графиком функции $\widehat{y} = \widehat{f}(x)$ – выборочной оценки функции $f(x)$, то можем указать два представления функции $\widehat{y}$:

$$\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x = \overline{y} + \widehat{b}_1(x - \overline{x}). \tag{1.24}$$

Коэффициент $\widehat{b}_1$ – *угловой коэффициент прямой регрессии*. Он показывает, на сколько единиц в среднем изменяется переменная y при увеличении переменной x на одну единицу.

Введем (выборочные) ковариации между всевозможными парами переменных x, y[1]:

$$k_{11} = \overline{x^2} - (\overline{x})^2,\ k_{1y} = \overline{xy} - \overline{x} \cdot \overline{y},\ k_{yy} = \overline{y^2} - (\overline{y})^2. \qquad (1.25)$$

Среди них k_{11}, k_{yy} (выборочные) дисперсии переменных x, y соответственно, причем $k_{11} > 0, k_{yy} > 0$.

Ковариации k_{11}, k_{1y}, k_{yy} – важные характеристики парной линейной регрессии. Угловой коэффициент $\widehat{b}_1$ записывается через ковариации следующим образом:

$$\widehat{b}_1 = \frac{k_{1y}}{k_{11}}\,. \qquad (1.26)$$

Из формулы (1.26) следует, что коэффициент $\widehat{b}_1$ пропорционален выборочному коэффициенту корреляции между переменными x, y

$$r_{\text{в}} = \frac{k_{1y}}{\sqrt{k_{11}} \cdot \sqrt{k_{yy}}}$$

и имеет его знак:

$$\widehat{b}_1 = r_{\text{в}} \cdot \frac{\sqrt{k_{yy}}}{\sqrt{k_{11}}}\,. \qquad (1.27)$$

Пример 1.1[2]. Рассматривается сеть магазинов. Изучается зависимость розничного товарооборота y_i, млн р. i-го магазина от числа его работников x_i. Статистические данные приведены в табл. 1.1.

Таблица 1.1

i	1	2	3	4	5	6	7	8
x_i	73	85	102	115	122	126	134	147
y_i	0,5	0,7	0.9	1,1	1,4	1,4	1,7	1,9

Требуется:

1) По расположению точек корреляционного поля на плоскости (x, y) убедиться, что целесообразно использование линейной регрессионной модели.

---

[1]Единица (1) в индексах обозначений ковариаций – символ регрессора $x = x_1$. Возможно, лучше было бы писать: k_{xx}, k_{xy}, но наши обозначения согласуются с символикой, принятой для общей регрессионной модели (см. подп. 1.2.3).

[2]См. работу В.А. Колемаева [11, с. 20].

2) Найти точечные оценки коэффициентов линейной регрессии. Построить на плоскости прямую регрессии.

Решение:

1) Характер корреляционного поля (рис. 1.2) позволяет применить линейную модель регрессии.

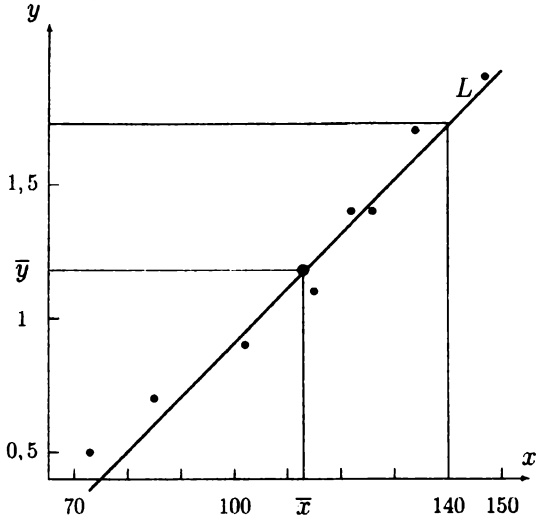


Рис. 1.2. Парная линейная регрессия. Пример 1.1

2) Вначале найдем средние величины $\overline{x}, \overline{y}, \overline{x^2}, \overline{y^2}, \overline{xy}$. Для этого заполняем расчетную табл. 1.2.

После этого последовательно вычислим

$$k_{11} \overset{(1.25)}{=} \overline{x^2} - (\overline{x})^2 = 13313,5 - 113^2 = 13313,5 - 12769 = 544,5;$$

$$k_{1y} = \overline{xy} - \overline{x} \cdot \overline{y} = 146,075 - 113 \cdot 1,2 = 146,075 - 135,6 = 10,475;$$

$$k_{yy} = \overline{y^2} - (\overline{y})^2 = 1,6475 - 1,2^2 = 0,2075;$$

$$\widehat{b}_1 = \frac{k_{1y}}{k_{11}} = \frac{10,475}{544,5} = 0,0192;$$

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x} = 1,2 - 0,0192 \cdot 113 = 1,2 - 2,173 = -0,973.$$

Поэтому уравнение прямой регрессии запишется в виде

$$L: \; y = -0,973 + 0,0192\,x.$$

Таблица 1.2

i	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	73	0,5	5329	0,25	36,5
2	85	0,7	7225	0,49	59,5
3	102	0,9	10404	0,81	91,8
4	115	1,1	13225	1,21	126,5
5	122	1,4	14884	1,96	170,8
6	126	1,4	15876	1,96	176,4
7	134	1,7	17956	2,89	227,8
8	147	1,9	21609	3,61	279,3
$\sum$	904	9,6	106508	13,18	1168,6
$\sum/n$	113	1,2	13313,5	1,6475	146,075
Средние	$\overline{x}$	$\overline{y}$	$\overline{x^2}$	$\overline{y^2}$	$\overline{xy}$

Коэффициент $\widehat{b}_1 = 0,0192$ означает, что при увеличении на одного человека численности работников магазина, входящего в сеть, товарооборот этого магазина увеличивается в среднем на 19200 р.

Статистический анализ построенной модели приведен ниже, в подп. 1.3.7.

1.2.2. Матричная форма модели линейной регрессии

Перейдем к рассмотрению множественной линейной регрессии ($m > 1$).

Исходная модель теперь запишется в виде

$$y = b_0 + b_1x_1 + \ldots + b_mx_m + \varepsilon, \tag{1.28}$$

а модель в наблюдениях (1.6), являющаяся для нас основной, в виде

$$y_i = b_0 \cdot 1 + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_m \cdot x_{im} + \varepsilon_i \quad (i \in \overline{1,n}). \tag{1.29}$$

Для исследования модели (1.29) представим ее более компактно – *в матричной форме.*

Введем вектор-столбец $\mathbf{b}$ коэффициентов регрессии размера

$(m+1) \times 1$ и вектор-столбец $\boldsymbol{\varepsilon}$ ошибок регрессии размера $n \times 1$:

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Особо важную роль среди всех коэффициентов регрессии играют коэффициенты *при регрессорах* $b_1, \dots, b_m$. Объединим их в вектор-столбец $\mathbf{b}_r$ размера $m \times 1$

$$\mathbf{b}_r = (b_1, \dots, b_m)^T.$$

Информацию о выборке (1.1) разместим в вектор-столбец $\mathbf{y}$ размера $n \times 1$ (переменная y) и в регрессионную матрицу $\mathbf{X}$ размера $n \times (m+1)$ (переменные $x_1, \dots x_m$):

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}.$$

Матрица $\mathbf{X}$ помимо выборочных значений регрессоров содержит первый столбец, состоящий из *единиц*.

В матричной форме модель (1.29) примет вид

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}. \tag{1.30}$$

Матричная модель (1.30), разумеется, может быть записана и в случае *парной* линейной регрессии ($m = 1$). Тогда вектор $\mathbf{b}$ состоит из двух элементов, а регрессионная матрица $\mathbf{X}$ – из двух столбцов:

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}.$$

Теперь регрессор x_1 обозначается через x, так что при записи элементов матрицы $\mathbf{X}$ нет необходимости в двойных индексах. Это обстоятельство необходимо учитывать при рассмотрении парной регрессии и в дальнейшем.

Для единообразия мы часто будем рассматривать матричную форму (1.30) модели линейной регрессии также и в случае $m = 1$.

1.2.3. Линейная регрессия (общий случай)

С помощью метода наименьших квадратов теперь найдем вектор точечных оценок

$$\widehat{\mathbf{b}} = (\widehat{b}_0, \widehat{b}_1, \ldots, \widehat{b}_m)^T$$

для неизвестного вектора $\mathbf{b}$ коэффициентов модели (1.30). Введем вектор остатков регрессии

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\mathbf{b}} = (e_1,\ e_2,\ \ldots,\ e_n)^T. \tag{1.31}$$

Применяя правила действий с матрицами (см. приложение), преобразуем сумму квадратов остатков регрессии:

$$\sum_{i=1}^{n} e_i^2 \overset{(\Pi 2)}{=} \mathbf{e}^T\mathbf{e} \overset{(1.31)}{=} (\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}) \overset{(\Pi 1)}{=} (\mathbf{y}^T - \widehat{\mathbf{b}}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}) =$$

$$= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y} + \widehat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\widehat{\mathbf{b}}.$$

Так как $\mathbf{y}^T\mathbf{X}\widehat{\mathbf{b}}$ — число, т.е. матрица размера 1×1, то эта матрица симметрична и совпадает с транспонированной к ней матрицей:

$$\mathbf{y}^T\mathbf{X}\widehat{\mathbf{b}} = (\mathbf{y}^T\mathbf{X}\widehat{\mathbf{b}})^T = \widehat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y}.$$

Поэтому, окончательно,

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - 2\widehat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y} + \widehat{\mathbf{b}}^T(\mathbf{X}^T\mathbf{X})\widehat{\mathbf{b}}. \tag{1.32}$$

Потребуем, чтобы вектор $\widehat{\mathbf{b}}$ был решением задачи минимизации

$$Q(\widehat{\mathbf{b}}) = \sum_{i=1}^{n} e_i^2 \to \min. \tag{1.33}$$

Убедимся, что функция $Q(\widehat{\mathbf{b}})$ имеет одну критическую точку, являющуюся точкой минимума.

Критические точки функции удовлетворяют матричному уравнению

$$\mathbf{grad}\, Q(\widehat{\mathbf{b}}) = \boldsymbol{\theta}, \tag{1.34}$$

где $\mathbf{grad}\, Q(\widehat{\mathbf{b}})$ – градиент функции $Q(\widehat{\mathbf{b}})$ (см. формулу (П10)), $\boldsymbol{\theta} = (0,\ 0,\ \ldots,\ 0)^T$ – нулевой вектор-столбец.

Учитывая свойства градиента и формулы (П11), (П12), находим

$$\mathbf{grad}\, Q(\widehat{\mathbf{b}}) = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X})\widehat{\mathbf{b}}.$$

Поэтому уравнение (1.34) преобразуется к линейному матричному уравнению относительно $\widehat{\mathbf{b}}$

$$(\mathbf{X}^T\mathbf{X})\,\widehat{\mathbf{b}} = \mathbf{X}^T\mathbf{y}. \tag{1.35}$$

Если *матрица* $\mathbf{X}^T\mathbf{X}$ *невырождена* или, что равносильно, если $\mathbf{X}$ – матрица полного (максимального) ранга $(m+1)$, т.е.

$$rank\ \mathbf{X} = m + 1, \tag{1.36}$$

то уравнение (1.35) имеет *единственное* решение

$$\widehat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{1.37}$$

являющееся критической точкой функции $Q(\widehat{\mathbf{b}})$. Проверяется, что найденная критическая точка – точка минимума, вследствие чего она и является единственным решением задачи (1.33).

Вычислив вектор $\widehat{\mathbf{b}}$, мы сможем по формуле (1.31) найти вектор остатков регрессии $\mathbf{e}$ и записать уравнение

$$\mathbf{y} = \mathbf{X}\,\widehat{\mathbf{b}} + \mathbf{e}. \tag{1.38}$$

Оно является оценкой исходной модели в наблюдениях (1.30) по выборке (1.1).

Введя вектор

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\mathbf{b}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T, \tag{1.39}$$

запишем уравнение (1.38) более кратко:

$$\mathbf{y} = \widehat{\mathbf{y}} + \mathbf{e}.$$

Вектор $\widehat{\mathbf{b}}$ – точечная МНК-оценка вектора $\mathbf{b}$. Координаты вектора $\widehat{\mathbf{b}}$ определяют в $(m+1)$-мерном пространстве $\mathbb{R}^{m+1}$ переменных $x_1, x_2, \ldots, x_m, y$ *гиперплоскость* T *линейной множественной регрессии*

$$T:\ y = \widehat{b}_0 + \widehat{b}_1 x_1 + \ldots + \widehat{b}_m x_m. \tag{1.40}$$

Поскольку ее коэффициенты являются решением задачи минимизации (1.33), она наиболее близка к корреляционному полю (1.1).

В случае двух регрессоров ($m = 2$) T – обычная *плоскость* в пространстве $\mathbb{R}^3$. Если же $m = 1$, то уравнение (1.40) принимает вид (1.12) и, значит, $T = L$ — *прямая* на плоскости $\mathbb{R}^2$.

В уравнении (1.40) основную роль играют коэффициенты $\widehat{b}_p$ ($p \in \overline{1,m}$) при регрессорах. Каждый коэффициент $\widehat{b}_p$ показывает, на сколько единиц в среднем изменится переменная y при увеличении регрессора x_p на одну единицу и при неизменной величине остальных регрессоров.

Правая часть уравнения (1.40) определяет функцию $\widehat{y} = \widehat{f}(x_1, \ldots, x_m)$, которая является оцененной функцией линейной регрессии (1.7):

$$\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x_1 + \ldots + \widehat{b}_m x_m. \tag{1.41}$$

Функцию (1.41) называют также функцией *линейной регрессии* y *на* $x_1, \ldots, x_m$.

Обсудим задачу *нахождения* вектора $\widehat{\mathbf{b}}$. Уравнение (1.35) есть матричная форма системы линейных неоднородных уравнений относительно коэффициентов $\widehat{b}_0, \widehat{b}_1, \ldots, \widehat{b}_m$. Осмыслим структуру этой системы. Найдем входящие в уравнение матрицы $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}^T\mathbf{y}$. Здесь $\mathbf{X}^T\mathbf{X}$ — квадратная симметрическая матрица порядка $(m+1)$, $\mathbf{X}^T\mathbf{y}$ – вектор-столбец размера $(m+1) \times 1$. Вычисляем[3]

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & & 1 \\ x_{11} & x_{21} & x_{31} & & x_{n1} \\ x_{12} & x_{22} & x_{32} & & x_{n2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{1m} & x_{2m} & x_{3m} & \ldots & x_{nm} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} =$$

$$= \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \ldots & \sum x_{i1}x_{im} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \ldots & \sum x_{i2}x_{im} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum x_{im} & \sum x_{i1}x_{im} & \sum x_{i2}x_{im} & \ldots & \sum x_{im}^2 \end{pmatrix}.$$

[3]Здесь и далее $\sum$ – краткое обозначение символики $\sum_{i=1}^{n}$.

Таким образом,

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \dots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & & \sum x_{i1}x_{im} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \dots & \sum x_{i2}x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_{im} & \sum x_{i1}x_{im} & \sum x_{i2}x_{im} & \dots & \sum x_{im}^2 \end{pmatrix}. \qquad (1.42)$$

Далее,

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1m} & x_{2m} & x_{3m} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \sum x_{i2}y_i \\ \dots \\ \sum x_{im}y_i \end{pmatrix}. \qquad (1.43)$$

Поэтому уравнение (1.35) преобразуется к виду системы *нормальных уравнений* – системы $(m+1)$ линейных неоднородных уравнений с $(m+1)$ неизвестными $\widehat{b}_0, \widehat{b}_1, \dots, \widehat{b}_m$:

$$\begin{cases} n\cdot\widehat{b}_0 + \sum x_{i1}\cdot\widehat{b}_1 + \sum x_{i2}\cdot\widehat{b}_2 + \dots + \\ \quad + \sum x_{im}\cdot\widehat{b}_m = \sum y_i, \\ \sum x_{i1}\cdot\widehat{b}_0 + \sum x_{i1}^2\cdot\widehat{b}_1 + \sum x_{i1}x_{i2}\cdot\widehat{b}_2 + \dots + \\ \quad + \sum x_{i1}x_{im}\cdot\widehat{b}_m = \sum x_{i1}y_i, \\ \sum x_{i2}\cdot\widehat{b}_0 + \sum x_{i1}x_{i2}\cdot\widehat{b}_1 + \sum x_{i2}^2\cdot\widehat{b}_2 + \dots + \\ \quad + \sum x_{i2}x_{im}\cdot\widehat{b}_m = \sum x_{i2}y_i, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \sum x_{im}\cdot\widehat{b}_0 + \sum x_{i1}x_{im}\cdot\widehat{b}_1 + \sum x_{i2}x_{im}\cdot\widehat{b}_2 + \dots + \\ \quad + \sum x_{im}^2\cdot\widehat{b}_m = \sum x_{im}y_i. \end{cases} \qquad (1.44)$$

Ее можно решить по формулам Крамера или методом Гаусса. Однако в ряде случаев систему удобнее решать *подстановкой*, сведя ее к системе m линейных уравнений относительно коэффициентов при регрессорах $\widehat{b}_1, \dots, \widehat{b}_m$. Предварительно введем средние величины $(p \in \overline{1,m},\ q \in \overline{1,m})$:

$$\overline{x}_p = \frac{\sum\limits_{i=1}^{n} x_{ip}}{n},\ \overline{x_p x_q} = \frac{\sum\limits_{i=1}^{n} x_{ip}x_{iq}}{n},\ \overline{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n},\ \overline{x_p y} = \frac{\sum\limits_{i=1}^{n} x_{ip}y_i}{n}. \qquad (1.45)$$

Разделив уравнения системы (1.44) на n, приведем ее к более простому виду

$$\begin{cases} \widehat{b}_0 + \overline{x}_1 \cdot \widehat{b}_1 + \overline{x}_2 \cdot \widehat{b}_2 + \ldots + \overline{x}_m \cdot \widehat{b}_m = \overline{y}, \\ \overline{x}_1 \cdot \widehat{b}_0 + \overline{x_1^2} \cdot \widehat{b}_1 + \overline{x_1 x_2} \cdot \widehat{b}_2 + \ldots + \overline{x_1 x_m} \cdot \widehat{b}_m = \overline{x_1 y}, \\ \overline{x}_2 \cdot \widehat{b}_0 + \overline{x_1 x_2} \cdot \widehat{b}_1 + \overline{x_2^2} \cdot \widehat{b}_2 + \ldots + \overline{x_2 x_m} \cdot \widehat{b}_m = \overline{x_2 y}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \overline{x}_m \cdot \widehat{b}_0 + \overline{x_1 x_m} \cdot \widehat{b}_1 + \overline{x_2 x_m} \cdot \widehat{b}_2 + \ldots + \overline{x_m^2} \cdot \widehat{b}_m = \overline{x_m y}. \end{cases} \tag{1.46}$$

Очевидно, что $\dfrac{1}{n}\mathbf{X}^T\mathbf{X}$ - матрица системы (1.46).

В случае парной линейной регрессии ($m = 1$) система (1.46) запишется в уже знакомом виде (1.20).

Из первого уравнения системы (1.46) выразим $\widehat{b}_0$ через остальные неизвестные:

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x}_1 - \ldots - \widehat{b}_m\overline{x}_m. \tag{1.47}$$

Подставив это выражение вместо $\widehat{b}_0$ в остальные уравнения системы, придем к системе m линейных уравнений относительно $\widehat{b}_1, \ldots, \widehat{b}_m$. Решение последней системы удобно записать в матричной форме. Проведем соответствующие выкладки. Прежде всего, исключая $\widehat{b}_0$ из всех уравнений системы (1.46), кроме первого, приходим к системе

$$\begin{cases} \overline{x}_1 \cdot (\overline{y} - \widehat{b}_1\overline{x}_1 - \ldots - \widehat{b}_m\overline{x}_m) + \overline{x_1^2} \cdot \widehat{b}_1 + \overline{x_1 x_2} \cdot \widehat{b}_2 + \\ + \ldots + \overline{x_1 x_m} \cdot \widehat{b}_m = \overline{x_1 y}, \\ \overline{x}_2 \cdot (\overline{y} - \widehat{b}_1\overline{x}_1 - \ldots - \widehat{b}_m\overline{x}_m) + \overline{x_1 x_2} \cdot \widehat{b}_1 + \overline{x_2^2} \cdot \widehat{b}_2 + \\ + \ldots + \overline{x_2 x_m} \cdot \widehat{b}_m = \overline{x_2 y}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \overline{x}_m \cdot (\overline{y} - \widehat{b}_1\overline{x}_1 - \ldots - \widehat{b}_m\overline{x}_m) + \overline{x_1 x_m} \cdot \widehat{b}_1 + \overline{x_2 x_m} \cdot \widehat{b}_2 + \\ + \ldots + \overline{x_m^2} \cdot \widehat{b}_m = \overline{x_m y}. \end{cases}$$

Преобразуем эту систему к виду системы линейных уравнений с

неизвестными $\widehat{b}_1, \ldots, \widehat{b}_m$:

$$\begin{cases} (\overline{x_1^2} - (\overline{x}_1)^2) \cdot \widehat{b}_1 + (\overline{x_1 x_2} - \overline{x}_1 \cdot \overline{x}_2) \cdot \widehat{b}_2 + \ldots + \\ +(\overline{x_1 x_m} - \overline{x}_1 \cdot \overline{x}_m) \cdot \widehat{b}_m = \overline{x_1 y} - \overline{x}_1 \cdot \overline{y}, \\ (\overline{x_1 x_2} - \overline{x}_1 \cdot \overline{x}_2) \cdot \widehat{b}_1 + (\overline{x_2^2} - (\overline{x}_2)^2) \cdot \widehat{b}_2 + \ldots + \\ +(\overline{x_2 x_m} - \overline{x}_2 \cdot \overline{x}_m) \cdot \widehat{b}_m = \overline{x_2 y} - \overline{x}_2 \cdot \overline{y}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (\overline{x_1 x_m} - \overline{x}_1 \cdot \overline{x}_m) \cdot \widehat{b}_1 + (\overline{x_2 x_m} - \overline{x}_2 \cdot \overline{x}_m) \cdot \widehat{b}_2 + \ldots + \\ +(\overline{x_m^2} - (\overline{x}_m)^2) \cdot \widehat{b}_m \overline{x_m y} - \overline{x}_m \cdot \overline{y}. \end{cases} \tag{1.48}$$

Для записи системы (1.48) в матричной форме введем (выборочные) ковариации между всевозможными парами переменных $x_1, \ldots, x_m, y \quad (p \in \overline{1,m},\ q \in \overline{1,m})$:

$$k_{pq} = \overline{x_p x_q} - \overline{x}_p \cdot \overline{x}_q,\ k_{py} = \overline{x_p y} - \overline{x}_p \cdot \overline{y},\ k_{yy} = \overline{y^2} - (\overline{y})^2. \tag{1.49}$$

Они являются элементами выборочной ковариационной матрицы переменных $x_1, \ldots, x_m, y$

$$\mathbf{K_в} = \begin{pmatrix} k_{11} & k_{12} & \ldots & k_{1m} & k_{1y} \\ k_{12} & k_{22} & \ldots & k_{2m} & k_{2y} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ k_{1m} & k_{2m} & \ldots & k_{mm} & k_{my} \\ k_{1y} & k_{2y} & \ldots & k_{my} & k_{yy} \end{pmatrix}$$

(см. формулу (П8)). Это – квадратная симметрическая матрица порядка $(m+1)$.

Рассмотрим следующие подматрицы матрицы $\mathbf{K_в}$: ковариационную матрицу регрессоров

$$\mathbf{K} = \begin{pmatrix} k_{11} & k_{12} & \ldots & k_{1m} \\ k_{12} & k_{22} & \ldots & k_{2m} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ k_{1m} & k_{2m} & \ldots & k_{mm} \end{pmatrix}, \tag{1.50}$$

вектор-столбец ковариаций между каждым регрессором и переменной y

$$\mathbf{k_y} = \left(k_{1y},\ k_{2y},\ \ldots,\ k_{my}\right)^T. \tag{1.51}$$

Тогда система (1.48) запишется в виде

$$\mathbf{K}\widehat{\mathbf{b}}_r = \mathbf{k_y}, \tag{1.52}$$

откуда при $\det \mathbf{K} \neq 0$ находим вектор

$$\widehat{\mathbf{b}}_r = \mathbf{K}^{-1}\mathbf{k_y}. \tag{1.53}$$

Он объединяет точечные оценки коэффициентов при регрессорах:

$$\widehat{\mathbf{b}}_r = (\widehat{b}_1, \ldots, \widehat{b}_m)^T.$$

Таким образом, возможны два способа нахождения точечных МНК-оценок коэффициентов линейной регрессии:

1) по формуле (1.37) оценивают сразу все коэффициенты;

2) применяя формулы (1.53), (1.47), вначале оценивают коэффициенты при регрессорах, а затем находят оценку свободного члена b_0.

Первый способ общеупотребителен в эконометрической литературе, элементы второго способа см. в работах Э. Кейна [10, с. 64]; Э. Фёрстера и Б. Рёнца [20, с.107].

В случае *одного* регрессора матрица $\mathbf{K}$ и векторы $\mathbf{b}_r, \mathbf{k_y}$ числа:

$$\mathbf{K} = k_{11},\ \mathbf{b}_r = b_1,\ \mathbf{k_y} = k_{1y}.$$

Поэтому формулы (1.47), (1.53) при $m = 1$ принимают привычный для парной регрессии вид (1.21), (1.26).

Выведем формулы МНК-оценок коэффициентов *трехмерной линейной регрессии* ($m = 2$). Тогда матричное уравнение (1.52) запишется в виде системы двух линейных уравнений

$$\begin{cases} k_{11}\,\widehat{b}_1 + k_{12}\,\widehat{b}_2 = k_{1y}, \\ k_{12}\,\widehat{b}_1 + k_{22}\,\widehat{b}_2 = k_{2y}. \end{cases}$$

Решая эту систему по правилу Крамера и учитывая равенство (1.47), приходим к формулам:

$$\widehat{b}_1 = \frac{k_{22}k_{1y} - k_{12}k_{2y}}{\Delta},\ \widehat{b}_2 = \frac{k_{11}k_{2y} - k_{12}k_{1y}}{\Delta}, \qquad \widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x}_1 - \widehat{b}_2\overline{x}_2 \quad (m = 2). \tag{1.54}$$

Здесь

$$\Delta = \det \mathbf{K} = k_{11} \cdot k_{22} - k_{12}^2.$$

Возвратимся к произвольной линейной регрессии. Ковариации (1.49) можно записать по-другому, связав их с суммами, включающими соответствующие разности $x_{ip} - \overline{x}_p$, $y_i - \overline{y}$ $(p \in \overline{1,m})$:

$$\sum_{i=1}^{n}(x_{ip} - \overline{x}_p)(x_{iq} - \overline{x}_q) = nk_{pq}\,,\ \sum_{i=1}^{n}(x_{ip} - \overline{x}_p)(y_i - \overline{y}) = nk_{py}\,,$$
$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = nk_{yy}\,. \tag{1.55}$$

Проверим, например, первую из этих формул:

$$\sum_{i=1}^{n}(x_{ip} - \overline{x}_p)(x_{iq} - \overline{x}_q) = \sum_{i=1}^{n} x_{ip}x_{iq} - \overline{x}_p \cdot \sum_{i=1}^{n} x_{iq} - \overline{x}_q \cdot \sum_{i=1}^{n} x_{ip} + n\overline{x}_p \cdot \overline{x}_q =$$

$$= n\overline{x_p x_q} - 2\overline{x}_p \cdot n\overline{x}_q + n\overline{x}_p \cdot \overline{x}_q = n(\overline{x_p x_q} - \overline{x}_p \cdot \overline{x}_q) \overset{(1.49)}{=} nk_{pq}\,.$$

Используя свойства определителей, убедимся, что матрицы систем (1.46), (1.48) имеют одинаковые определители:

$$\det\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right) = \det\mathbf{K}. \tag{1.56}$$

Для этого обнулим все элементы первой строки определителя $\det\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)$, кроме 1, а затем разложим определитель по первой строке:

$$\det\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right) = \begin{vmatrix} 1 & \overline{x}_1 & \overline{x}_2 & \dots & \overline{x}_m \\ \overline{x}_1 & \overline{x_1^2} & \overline{x_1x_2} & \dots & \overline{x_1x_m} \\ \overline{x}_2 & \overline{x_1x_2} & \overline{x_2^2} & \dots & \overline{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ \overline{x}_m & \overline{x_1x_m} & \overline{x_2x_m} & \dots & \overline{x_m^2} \end{vmatrix} =$$

$$= \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ \overline{x}_1 & \overline{x_1^2} - (\overline{x}_1)^2 & \overline{x_1x_2} - \overline{x}_1 \cdot \overline{x}_2 & \dots & \overline{x_1x_m} - \overline{x}_1 \cdot \overline{x}_m \\ \overline{x}_2 & \overline{x_1x_2} - \overline{x}_1 \cdot \overline{x}_2 & \overline{x_2^2} - (\overline{x}_2)^2 & \dots & \overline{x_2x_m} - \overline{x}_2 \cdot \overline{x}_m \\ \dots & \dots & \dots & \dots & \dots \\ \overline{x}_m & \overline{x_1x_m} - \overline{x}_1 \cdot \overline{x}_m & \overline{x_2x_m} - \overline{x}_2 \cdot \overline{x}_m & \dots & \overline{x_m^2} - (\overline{x}_m)^2 \end{vmatrix} =$$

$$
= \begin{vmatrix} \overline{x_1^2} - (\overline{x}_1)^2 & \overline{x_1 x_2} - \overline{x}_1 \cdot \overline{x}_2 & \ldots & \overline{x_1 x_m} - \overline{x}_1 \cdot \overline{x}_m \\ \overline{x_1 x_2} - \overline{x}_1 \cdot \overline{x}_2 & \overline{x_2^2} - (\overline{x}_2)^2 & \ldots & \overline{x_2 x_m} - \overline{x}_2 \cdot \overline{x}_m \\ \ldots & \ldots & \ldots & \ldots \\ \overline{x_1 x_m} - \overline{x}_1 \cdot \overline{x}_m & \overline{x_2 x_m} - \overline{x}_2 \cdot \overline{x}_m & \ldots & \overline{x_m^2} - (\overline{x}_m)^2 \end{vmatrix} \overset{(1.49)}{=} \det \mathbf{K}.
$$

Свойства МНК-оценок

Укажем особенности МНК, напрямую не связанные со статистической природой оцениваемой линейной регрессионной модели.

1. *Вектор* $\widehat{\mathbf{b}}$ *линейная (относительно* $\mathbf{y}$*) оценка неизвестного вектора* $\mathbf{b}$.

Доказательство. Так как согласно формуле (1.37) вектор $\widehat{\mathbf{b}}$ есть результат умножения вектора наблюдений

$$
\mathbf{y} = (y_1, \ldots, y_n)^T
$$

слева на матрицу $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, то координаты вектора $\widehat{\mathbf{b}}$ являются линейными комбинациями величин $y_1, \ldots, y_n$.

2. *Гиперплоскость линейной регрессии проходит через точку*

$$
(\overline{x}_1, \ldots, \overline{x}_m, \overline{y}).
$$

Доказательство следует из уравнения линейной регрессии (1.40) и из формулы (1.47).

3. *Сумма остатков регрессии равна нулю:*

$$
\sum_{i=1}^{n} e_i = 0. \tag{1.57}
$$

Доказательство. Вначале убедимся, что произведение матрицы $\mathbf{X}^T$ на вектор $\mathbf{e}$ является нулевым вектором:

$$
\mathbf{X}^T\mathbf{e} = \boldsymbol{\theta}. \tag{1.58}
$$

Действительно,

$$
\mathbf{X}^T\mathbf{e} \overset{(1.31)}{=} \mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\widehat{\mathbf{b}} \overset{(1.35)}{=} \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} = \boldsymbol{\theta}.
$$

Запишем равенство (1.58) подробнее:

$$
\begin{pmatrix} 1 & & 1 \\ x_{11} & \ldots & x_{n1} \\ \ldots & \ldots & \ldots \\ x_{1m} & \ldots & x_{nm} \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{pmatrix}; \quad \begin{pmatrix} e_1 + \ldots + e_n \\ x_{11}e_1 + \ldots + x_{n1}e_n \\ \ldots \\ x_{1m}e_1 + \ldots + x_{nm}e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{pmatrix}.
$$

Приравнивая *первые* координаты вектор-столбцов, стоящих в левой и правой частях последнего равенства, приходим к требуемому равенству (1.57).

Из второго свойства МНК-оценок вытекают новые представления гиперплоскости T линейной регрессии и функции $\widehat{y}$ линейной регрессии y на $x_1, \ldots, x_m$

$$T: \; y = \overline{y} + \widehat{b}_1(x_1 - \overline{x}_1) + \ldots + \widehat{b}_m(x_m - \overline{x}_m); \tag{1.59}$$

$$\widehat{y} = \overline{y} + \widehat{b}_1(x_1 - \overline{x}_1) + \ldots + \widehat{b}_m(x_m - \overline{x}_m) \tag{1.60}$$

(сравните с формулами (1.40), (1.41)).

Пример 1.2. Анализируется объем инвестиций y (тыс. у.е.) некоторой фирмы за 10 лет. Предполагается, что переменная y зависит от дохода фирмы x_1 (тыс. у.е.) и от величины процентной ставки x_2 (%). Статистические данные приведены в табл. 1.3 .

Таблица 1.3

x_{1i}	100	110	140	150	160	160	180	200	230	250
x_{2i}	2	2	3	2	3	4	4	3	4	5
y_i	20	25	30	30	35	38	40	38	44	50

Требуется:

1) Найти точечные оценки коэффициентов линейной регрессии y на x_1, x_2. Составить уравнение плоскости регрессии.

Решение:

1) Для нахождения средних величин $\overline{x_1}, \overline{x_2}, \overline{y}, \overline{x_1^2}, \overline{x_2^2}, \overline{y^2}, \overline{x_1x_2}, \overline{x_1y}, \overline{x_2y}$ заполним табл. 1.4.

После этого вычислим выборочные ковариации между переменными:

$$k_{11} = \overline{x_1^2} - (\overline{x}_1)^2 = 30320 - 168^2 = 30320 - 28224 = 2096;$$

$$k_{22} = \overline{x_2^2} - (\overline{x}_2)^2 = 11,2 - 3,2^2 = 11,2 - 10,24 = 0,96;$$

$$k_{12} = \overline{x_1x_2} - \overline{x}_1 \cdot \overline{x}_2 = 575 - 168 \cdot 3,2 = 575 - 537,6 = 37,4;$$

$$k_{1y} = \overline{x_1y} - \overline{x}_1 \cdot \overline{y} = 6255 - 168 \cdot 35 = 6255 - 5880 = 375;$$

$$k_{2y} = \overline{x_2y} - \overline{x}_2 \cdot \overline{y} = 119,7 - 3,2 \cdot 35 = 119.7 - 112 = 7,7;$$

$$k_{yy} = \overline{y^2} - (\overline{y})^2 = 1297,4 - 35^2 = 1297,4 - 1225 = 72,4.$$

Таблица 1.4

i	x_{1i}	x_{2i}	y_i	x_{1i}^2	x_{2i}^2	y_i^2	$x_{1i}x_{2i}$	$x_{1i}y_i$	$x_{2i}y_i$
1	100	2	20	10000	4	400	200	2000	40
2	110	2	25	12100	4	625	220	2750	50
3	140	3	30	19600	9	900	420	4200	90
4	150	2	30	22500	4	900	300	4500	60
5	160	3	35	25600	9	1225	480	5600	105
6	160	4	38	25600	16	1444	640	6080	152
7	180	4	40	32400	16	1600	720	7200	160
8	200	3	38	40000	9	1444	600	7600	114
9	230	4	44	52900	16	1936	920	10120	176
10	250	5	50	62500	25	2500	1250	12500	250
$\sum$	1680	32	350	303200	112	12974	5750	62550	1197
$\sum/n$	168	3,2	35	30320	11,2	1297,4	575	6255	119,7
Средние	$\overline{x}_1$	$\overline{x}_2$	$\overline{y}$	$\overline{x_1^2}$	$\overline{x_2^2}$	$\overline{y^2}$	$\overline{x_1x_2}$	$\overline{x_1y}$	$\overline{x_2y}$

Находим

$$\Delta = \det \mathbf{K} = k_{11}k_{22} - k_{12}^2 = 2096 \cdot 0,96 - 37,4^2 = 2012,16 - 1398,76 = 613,4.$$

Сейчас сможем вычислить точечные оценки коэффициентов регрессии:

$$\widehat{b}_1 \overset{(1.54)}{=} \frac{k_{22}k_{1y} - k_{12}k_{2y}}{\Delta} = \frac{0,96 \cdot 375 - 37,4 \cdot 7,7}{613,4} = \frac{360 - 287,98}{613.4} = 0,117;$$

$$\widehat{b}_2 = \frac{k_{11}k_{2y} - k_{12}k_{1y}}{\Delta} = \frac{2096 \cdot 7,7 - 37,4 \cdot 375}{613,4} = 3,447;$$

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x}_1 - \widehat{b}_2\overline{x}_2 = 35 - 0,117 \cdot 168 - 3,447 \cdot 3,2 = 4,314.$$

Мы нашли уравнение регрессии y на x_1, x_2 (уравнение плоскости регрессии T в пространстве переменных x_1, x_2, y)

$$T: \quad y = 4,314 + 0.117x_1 + 3,447x_2.$$

Укажем экономический смысл коэффициентов при регрессорах . Так как $\widehat{b}_1 = 0,117$, то при увеличении дохода фирмы на 1 единицу, т.е. на

1 тыс. у. е., объем инвестиций в среднем возрастет на 117 у.е. Аналогично при возрастании процентной ставки на 1 % объем инвестиций в среднем возрастет на 3447 у.е.

Исследование построенной модели см. в подп. 1.3.7.

1.3. Статистический анализ классической линейной регрессионной модели

Переходим к изучению *статистических* свойств линейной регрессионной модели. Для этого необходимо наложить определенные условия на статистическую природу модели. В данном пункте мы рассмотрим простой, но практически важный класс моделей – *классические линейные регрессионные модели.*

1.3.1. Условия Гаусса–Маркова. Статистические свойства точечных оценок коэффициентов регрессии

Потребуем, чтобы модель (1.30) удовлетворяла следующим условиям, которые принято называть *условиями Гаусса–Маркова:*

1^0. Все ошибки ε_i – *случайные* величины, а регрессоры – *неслучайные* (детерминированные) переменные. Иначе говоря, $\boldsymbol{\varepsilon}$ – случайный вектор, $\mathbf{X}$ – неслучайная матрица.

2^0. Математическое ожидание каждой ошибки ε_i равно нулю: $M(\varepsilon_i) = 0$. Это равносильно тому, что $M(\boldsymbol{\varepsilon}) = \boldsymbol{\theta}$.

3^0. (*Условие гомоскедастичности – равноизменчивости ошибок.*) Дисперсии всех ошибок одинаковы: $D(\varepsilon_i) = \sigma^2 = const$.

4^0. (*Условие отсутствия автокорреляции.*) Различные ошибки ε_i некоррелированы между собой: $cov(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$. Здесь $cov(\varepsilon_i, \varepsilon_j)$ – ковариация между случайными величинами ε_i, ε_j.

5^0. Ошибки ε_i – нормально распределенные случайные величины: $\varepsilon_i \sim N(0, \sigma^2)$.

6^0. $rank\, \mathbf{X} = m + 1$.

Модель (1.30) с условиями $1^0 - 6^0$ называют *классической (нормальной) линейной регрессионной моделью.*

Условие 6^0 уже возникло ранее (см. формулу (1.36)) как условие единственности решения линейного уравнения (1.35). Оно означает, что столбцы регрессионной матрицы $\mathbf{X}$ линейно независимы. Иначе говоря, выборка не выявила линейной функциональной связи между регрессорами; кроме того, наблюдения за каждым регрессором не все одинаковы[4].

Объединим условия $3^0 - 4^0$, записав их через матрицу ковариаций $\mathbf{K}(\boldsymbol{\varepsilon})$ случайного вектора ошибок регрессии $\boldsymbol{\varepsilon}$. Полагая в общей формуле (П7) $\mathbf{a} = \boldsymbol{\varepsilon}$ и учитывая равенства

$$D(\varepsilon_i) \overset{3^0}{=} \sigma^2, \ cov(\varepsilon_i, \varepsilon_j) \overset{4^0}{=} 0 \ (i \neq j),$$

убеждаемся, что матрица $\mathbf{K}(\boldsymbol{\varepsilon})$ диагональна:

$$\mathbf{K}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma^2 & 0 & & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{E}.$$

Окончательно

$$3^0.4^0 \quad \Leftrightarrow \quad \mathbf{K}(\boldsymbol{\varepsilon}) = M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{E}. \tag{1.61}$$

Назовем параметры классической линейной регрессионной модели: коэффициенты регрессии $b_0, b_1, \dots, b_m$, дисперсия ошибок регрессии σ^2. Все они требуют оценивания по выборке. Точечные МНК-оценки коэффициентов регрессии мы уже получили. Точечная оценка параметра σ^2 дается в подп. 1.3.2.

Замечание 1.1. Поскольку условие 6^0 равносильно требованию невырожденности матрицы $\mathbf{X}^T\mathbf{X}$, то в силу равенства (1.56) оно представимо в виде

$$\det \mathbf{K} \neq 0.$$

Замечание 1.2. Условия Гаусса–Маркова можно записать через наблюдения y_i за зависимой переменной y. При этом нужно учесть детерминированность величин x_{il} – наблюдений за регрессорами. Нам понадобятся такие свойства y_i :

$$y_i \text{ – нормально распределенные случайные величины,} \tag{1.62}$$

[4]Условие 6^0 означает отсутствие строгой мультиколлинеарности (см. подп. 1.5.4).

$$D(y_i) = \sigma^2 = const. \tag{1.63}$$

В дальнейшем (см. п. 1.5) рассмотрим *обобщения* классической модели, где некоторые из требований 1^0 – 6^0 будут ослаблены.

Из курса математической статистики известно, что *качество* точечных оценок параметров случайной величины определяется свойствами *несмещенности, эффективности, состоятельности.* Напомним соответствующие определения. Пусть b – оцениваемый параметр (возможно, векторный), $\widehat{b}$ – его точечная оценка.

Определение 1.1. Оценка $\widehat{b}$ параметра b называется несмещенной, если

$$M(\widehat{b}) = b.$$

Это важное свойство означает отсутствие систематических ошибок оценивания.

Определение 1.2. Оценка $\widehat{b}$ называется эффективной (в некотором классе оценок), если она самая точная в этом классе, т.е. имеет *наименьшую дисперсию.*

Определение 1.3. Оценка $\widehat{b}$ называется состоятельной, если она сходится по вероятности к истинному значению b при $n \to \infty$.

Теорема Гаусса–Маркова. *Если линейная регрессионная модель (1.30) удовлетворяет условиям* 1^0 – 6^0, *то точечная оценка (1.37), полученная МНК, является несмещенной и эффективной (в классе всех несмещенных линейных оценок) оценкой вектора* $\mathbf{b}$.

Докажем лишь несмещенность оценки (1.37), т.е. что

$$M(\widehat{\mathbf{b}}) = \mathbf{b}.$$

Вначале преобразуем

$$\widehat{\mathbf{b}} \overset{(1.37)}{=} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \overset{(1.30)}{=} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{b}) +$$

$$+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} = \left[\, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{E}\mathbf{b} = \mathbf{b} \,\right] =$$

$$= \mathbf{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}.$$

Здесь выражение, заключенное в квадратные скобки, – комментарий, поясняющий выкладки.

Таким образом,

$$\widehat{\mathbf{b}} = \mathbf{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}. \tag{1.64}$$

Поэтому

$$M(\widehat{\mathbf{b}}) \overset{(1.64)}{=} M(\mathbf{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}) = M(\mathbf{b}) + M((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}) \overset{1^0}{=}$$

$$= \mathbf{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T M(\boldsymbol{\varepsilon}) \overset{2^0}{=} \mathbf{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\theta} = \mathbf{b} + \boldsymbol{\theta} = \mathbf{b}.$$

Замечание 1.3. При весьма слабых ограничениях на матрицу $\mathbf{X}$ МНК-оценка (1.37) вектора $\mathbf{b}$ является также и *состоятельной*[5].

1.3.2. Стандартные ошибки регрессии и коэффициентов регрессии

Переходим к описанию процедур, соответствующих этапу верификации эконометрического моделирования. Они связаны с проверкой статистического качества модели и оценкой ее точности.

Точность линейной регрессионной модели выражается через 1) стандартную ошибку регрессии s и 2) стандартные ошибки коэффициентов регрессии s_p ($p \in \overline{0,m}$). Введем эти характеристики.

1. Основной характеристикой точности классической линейной регрессионной модели служит дисперсия σ^2 ошибок регрессии ε_i (см. условие 3^0 Гаусса Маркова). Так как она обычно неизвестна, то приходится заменять ее точечной выборочной оценкой. Поскольку остаток регрессии e_i – точечная оценка ошибки ε_i, то естественно предположить, что искомая оценка содержит остатки.

Доказывается [12, с. 95 – 97], что несмещенной оценкой дисперсии σ^2 является выборочная остаточная дисперсия

$$s^2 = \frac{\sum\limits_{i=1}^{n} e_i^2}{n - m - 1}. \tag{1.65}$$

Для получения несмещенной оценки сумму квадратов остатков e_i мы делим на *число степеней свободы*, равное разности между числом наблюдений n и числом связей, ограничивающих свободу их изменения; число связей $(m+1)$ совпадает с числом ограничений – числом уравнений в системе (1.46).

---

[5]Укажем такое условие состоятельности: $\lambda_{\min} \to +\infty$ при $n \to \infty$ (здесь $\lambda_{\min}$ – наименьшее собственное значение матрицы $\mathbf{X}^T\mathbf{X}$) [1, с. 641].

Величина

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-m-1}} \tag{1.66}$$

называется *стандартной ошибкой линейной регрессии.*

2. Мерой точности коэффициентов регрессии $\widehat{b}_p$ ($p \in \overline{0,m}$) являются их дисперсии $D(\widehat{b}_p)$. Эти дисперсии – диагональные элементы ковариационной матрицы вектора коэффициентов $\widehat{\mathbf{b}}$

$$\mathbf{K}(\widehat{\mathbf{b}}) = M((\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\mathbf{b}} - \mathbf{b})^T)$$

(см. формулу (П7)). Найдем матрицу $\mathbf{K}(\widehat{\mathbf{b}})$. Убедимся, что она лишь множителем отличается от матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$\mathbf{K}(\widehat{\mathbf{b}}) = \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}. \tag{1.67}$$

Для этого преобразуем

$$\mathbf{K}(\widehat{\mathbf{b}}) = M((\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\mathbf{b}} - \mathbf{b})^T) \overset{(1.64)}{=} M((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon})^T) =$$

$$= M((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \overset{1^0}{=} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} =$$

$$= \left[M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) \overset{(1.61)}{=} \sigma^2\mathbf{E} \right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{E}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} =$$

$$= \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}.$$

Введем более краткую символику:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{Z} = (z_{pq}). \tag{1.68}$$

Здесь $p \in \overline{0,m}$, $q \in \overline{0,m}$. Тогда формула (1.67) запишется так:

$$\mathbf{K}(\widehat{\mathbf{b}}) = \sigma^2\mathbf{Z}. \tag{1.69}$$

Поэтому

$$D(\widehat{b}_p) = \sigma^2 z_{pp}, \tag{1.70}$$

где z_{pp} – p-й диагональный элемент матрицы $\mathbf{Z}$.

Часто удобнее использовать не дисперсию, а среднее квадратическое (стандартное) отклонение

$$\sqrt{D(\widehat{b}_p)} = \sigma \cdot \sqrt{z_{pp}}. \tag{1.71}$$

В практической работе вместо величины σ применяют ее несмещенную оценку s. Производя такую замену в правой части формулы (1.71), приходим к *стандартной ошибке коэффициента* $\widehat{b}_p$:

$$s_p = s \cdot \sqrt{z_{pp}} \quad (p \in \overline{0,m}). \tag{1.72}$$

1.3.3. Коэффициент детерминации. Оценка качества линейной регрессионной модели в целом

Проверка *статистического качества* регрессионной модели обычно включает следующие элементы:

- оценка качества модели в целом;
- проверка статистической значимости коэффициентов $\widehat{b}_l$ при регрессорах (в случае множественной регрессии);
- проверка предпосылок регрессии (условий Гаусса–Маркова).

В этом подпункте остановимся на первой задаче, в подп. 1.3.4 – на второй задаче. Некоторые подходы к решению последней задачи см. в подп. 1.5.2 1.5.3.

Проведем оценку качества классической нормальной регрессионной модели (1.30) *в целом*, иначе говоря, проверку *соответствия модели наблюдениям*. Такая проверка проводится на основе дисперсионного анализа, изучаемого в курсе математической статистики, и предполагает применение *коэффициента детерминации* R^2. Введем это понятие. Рассмотрим зависимую переменную y. Сравним ее выборочные значения y_i и объясняемые регрессией значения $\widehat{y}_i$ с выборочным средним $\overline{y}$.

Рассмотрим *общую вариацию* (разброс) наблюдаемых значений переменной y относительно выборочного среднего $\overline{y}$ – сумму квадратов разностей $(y_i - \overline{y})$

$$S = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{1.73}$$

Введем также вариацию, *объясняемую регрессией*

$$\widehat{S} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2, \tag{1.74}$$

и *остаточную* вариацию

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2. \tag{1.75}$$

Докажем, что эти вариации связаны формулой

$$S = \widehat{S} + S_e. \tag{1.76}$$

Преобразуем

$$S = \sum (y_i - \overline{y} - \widehat{y}_i + \widehat{y}_i)^2 = \sum ((\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i))^2 = [\, y_i - \widehat{y}_i = e_i \,] =$$

$$= \sum ((\widehat{y}_i - \overline{y}) + e_i)^2 = \sum (\widehat{y}_i - \overline{y})^2 + \sum e_i^2 + 2 \sum (\widehat{y}_i - \overline{y}) e_i.$$

Для проверки формулы (1.76) достаточно убедиться, что

$$\sum (\widehat{y}_i - \overline{y}) e_i = 0.$$

Учитывая свойства МНК-оценок (см. подп. 1.2.3), преобразуем

$$\sum (\widehat{y}_i - \overline{y}) e_i = \sum \widehat{y}_i e_i - \sum \overline{y} e_i = \sum \widehat{y}_i e_i - \overline{y} \sum e_i \overset{(1.57)}{=} \sum \widehat{y}_i e_i =$$

$$= \widehat{\mathbf{y}}^T \mathbf{e} \overset{(1.39)}{=} (\mathbf{X}\widehat{\mathbf{b}})^T \mathbf{e} = \widehat{\mathbf{b}}^T \mathbf{X}^T \mathbf{e} = \widehat{\mathbf{b}}^T (\mathbf{X}^T \mathbf{e}) \overset{(1.58)}{=} \widehat{\mathbf{b}}^T \boldsymbol{\theta} = 0.$$

Формула (1.76) доказана.

Величины $S, \widehat{S}, S_e$ неотрицательны, причем

$$\widehat{S} \leqslant S, \; S_e \leqslant S.$$

Коэффициент детерминации определяется формулой

$$R^2 = \frac{\widehat{S}}{S}. \tag{1.77}$$

Он указывает *часть (долю) общей вариации переменной y, объясняемую регрессией.*

Коэффициент детерминации вычисляется с помощью формул (1.73), (1.74), (1.77). Однако в ряде случаев его удобно находить по другой формуле, не требующей вычисления величин $\widehat{y}_i, e_i$. Для ее вывода преобразуем сумму квадратов остатков S_e, учитывая, что вектор $\widehat{\mathbf{b}}$ определен из условия ее минимизации (см. (1.33)). По формуле (1.47) исключим первую координату вектора $\widehat{\mathbf{b}}$ и перейдем к вектору $\widehat{\mathbf{b}}_r$:

$$S_e = \sum e_i^2 \overset{(1.32)}{=} \mathbf{y}^T \mathbf{y} - 2\widehat{\mathbf{b}}^T \mathbf{X}^T \mathbf{y} + \widehat{\mathbf{b}}^T (\mathbf{X}^T \mathbf{X}) \widehat{\mathbf{b}} \overset{(1.35)}{=} \mathbf{y}^T \mathbf{y} - 2\widehat{\mathbf{b}}^T \mathbf{X}^T \mathbf{y} +$$

$$+ \widehat{\mathbf{b}}^T \mathbf{X}^T \mathbf{y} = \mathbf{y}^T \mathbf{y} - \widehat{\mathbf{b}}^T \mathbf{X}^T \mathbf{y} = \left[\mathbf{y}^T \mathbf{y} = \sum y_i^2 = n\overline{y^2} \right] = n\overline{y^2} -$$

$$-\widehat{\mathbf{b}}^T(\mathbf{X}^T\mathbf{y}) \overset{(1.43)}{=} n\overline{y^2} - \widehat{b}_0 \cdot n\overline{y} - \widehat{b}_1 \cdot n\overline{x_1 y} - \ldots - \widehat{b}_m \cdot n\overline{x_m y} \overset{(1.47)}{=} n \cdot (\overline{y^2} -$$

$$-\overline{y} \cdot (\overline{y} - \widehat{b}_1 \cdot \overline{x}_1 - \ldots - \widehat{b}_m \overline{x}_m) - \widehat{b}_1 \cdot \overline{x_1 y} - \ldots - \widehat{b}_m \cdot \overline{x_m y}) =$$

$$= n \cdot (\overline{y^2} - (\overline{y})^2 - (\overline{x_1 y} - \overline{x}_1 \cdot \overline{y}) \cdot \widehat{b}_1 - \ldots - (\overline{x_m y} - \overline{x}_m \cdot \overline{y}) \cdot \widehat{b}_m) \overset{(1.49)}{=} n(k_{yy} -$$

$$-\widehat{b}_1 \cdot k_{1y} - \ldots - \widehat{b}_m \cdot k_{my}) = n(k_{yy} - \widehat{\mathbf{b}}_r^T \mathbf{k}_\mathbf{y}).$$

Теперь переходим к R^2 :

$$R^2 = 1 - \frac{S_e}{S} = 1 - \frac{S_e}{\sum (y_i - \overline{y})^2} \overset{(1.55)}{=} 1 - \frac{S_e}{nk_{yy}} =$$

$$= 1 - \frac{n(k_{yy} - \widehat{\mathbf{b}}_r^T \mathbf{k}_\mathbf{y})}{nk_{yy}} = \frac{\widehat{\mathbf{b}}_r^T \mathbf{k}_\mathbf{y}}{k_{yy}} \overset{(1.51)}{=} \frac{\sum\limits_{i=1}^{m} \widehat{b}_i \, k_{iy}}{k_{yy}} .$$

Окончательно

$$R^2 = \frac{\widehat{\mathbf{b}}_r^T \mathbf{k}_\mathbf{y}}{k_{yy}} = \frac{\sum\limits_{i=1}^{m} \widehat{b}_i \, k_{iy}}{k_{yy}} . \tag{1.78}$$

Из этой общей формулы вытекает следующая формула нахождения коэффициента детерминации в случае *трехмерной* линейной регрессии:

$$R^2 = \frac{\widehat{b}_1 k_{1y} + \widehat{b}_2 k_{2y}}{k_{yy}} \quad (m = 2). \tag{1.79}$$

Выведем удобную формулу, выражающую через коэффициент детерминации стандартную ошибку регрессии s (1.66). Так как

$$R^2 = 1 - \frac{S_e}{S} ,$$

то

$$S_e = S(1 - R^2) = nk_{yy}(1 - R^2),$$

поэтому

$$s = \sqrt{\frac{nk_{yy}(1 - R^2)}{n - m - 1}} . \tag{1.80}$$

В случае парной линейной регрессии ($m = 1$) формулы (1.78), (1.80) упрощаются. Тогда

$$\widehat{\mathbf{b}}_r = \widehat{b}_1, \ \mathbf{k}_\mathbf{y} = k_{1y},$$

отсюда

$$R^2 = \frac{\widehat{b}_1 \cdot k_{1y}}{k_{yy}} \overset{(1.26)}{=} \frac{k_{1y}^2}{k_{11}k_{yy}} = r_{\text{в}}^2 .$$

Далее преобразуем

$$1 - R^2 = 1 - \frac{\widehat{b}_1 \cdot k_{1y}}{k_{yy}} = \frac{k_{yy} - \widehat{b}_1 \cdot k_{1y}}{k_{yy}}.$$

Но тогда

$$s \overset{(1.80)}{=} \sqrt{\frac{nk_{yy}(1-R^2)}{n-2}} = \sqrt{\frac{nk_{yy}(k_{yy} - \widehat{b}_1 \cdot k_{1y})}{(n-2)k_{yy}}} = \sqrt{\frac{n(k_{yy} - \widehat{b}_1 \cdot k_{1y})}{n-2}} .$$

Таким образом,

$$R^2 = r_{\text{в}}^2 = \frac{k_{1y}^2}{k_{11}k_{yy}},\ s = \sqrt{\frac{n(k_{yy} - \widehat{b}_1 \cdot k_{1y})}{n-2}} \quad (m = 1). \tag{1.81}$$

Коэффициент детерминации (в любой классической линейной регрессионной модели) удовлетворяет неравенствам

$$0 \leqslant R^2 \leqslant 1.$$

Если $R^2 = 1$, то $e_i = 0$, $i \in \overline{1,n}$. Значит, все точки корреляционного поля лежат на гиперплоскости регрессии (1.40). Поэтому значения R^2, близкие к 1, являются, вообще говоря, индикатором высокого качества построенной регрессионной модели.

Формальная проверка соответствия модели наблюдениям проводится в рамках теории статистических гипотез. Выдвигается основная гипотеза H_0 о *равенстве нулю истинных коэффициентов при регрессорах*

$$H_0 :\ b_1 = b_2 = \ldots = b_m = 0.$$

Она означает, что функция линейной регрессии (1.7) принимает очень простой вид $y = b_0$ и, значит, линейная связь между переменными $x_1, \ldots, x_m, y$ отсутствует. Для проверки гипотезы используется F-статистика

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} . \tag{1.82}$$

Применение F-статистики опирается на 5^0 условие Гаусса–Маркова о нормальном распределении ошибок регрессии ε_i. Наблюдения y_i а, значит, и разности $y_i - \overline{y}$ обладают этим же свойством (см. (1.62)). Исходя из этого, можно доказать [14, с. 79-80], что в случае справедливости гипотезы H_0 F-статистика имеет распределение Фишера $F(k_1, k_2)$ со степенями свободы $k_1 = m,\ k_2 = n - m - 1$.

Применяя формулу (1.82), вычисляют наблюдаемое по модели $F_{\text{набл}}$ значение F-статистики. Выбирают уровень значимости α, равный вероятности того, что будет отвергнута истинная гипотеза H_0. По статистическим таблицам находят квантиль[6] распределения Фишера порядка $(1-\alpha)$: $\quad F_{\text{крит}} = F_{1-\alpha}(m, n-m-1)$.

Используется следующее разрешающее правило:

1) Если

$$F_{\text{набл}} > F_{\text{крит}}\ ,$$

то гипотеза H_0 отвергается, что свидетельствует о согласованности (на уровне значимости α) линейной регрессии результатам наблюдений.

2) Если же $F_{\text{набл}} \leqslant F_{\text{крит}}$, то гипотеза сохраняется. Скорее всего, требуется скорректировать изучаемую линейную модель.

Замечание 1.4. Следует с осторожностью применять коэффициент детерминации для выбора между *несколькими* регрессионными моделями одного и того же экономического объекта. Дело в том, что коэффициент R^2 возрастает при добавлении новых регрессоров, изменяется даже при простых преобразованиях переменных [14, с. 75–76]. Поэтому сравнивать посредством коэффициента R^2 целесообразно модели с одинаковыми регрессорами.

1.3.4. Оценка значимости коэффициентов множественной линейной регрессии

Специфической задачей *множественного* регрессионного анализа является *оценка влияния различных регрессоров* x_l $(l \in \overline{1,m})$ на модель с возможной корректировкой набора регрессоров.

---

[6] В книгах по математической статистике и эконометрике содержатся различные виды таблиц статистических распределений (Фишера, Стьюдента). Вместо квантилей могут указываться *критические* или *процентные точки* распределений.

Вначале рассмотрим *более общую* задачу о проверке гипотез об истинных коэффициентах регрессии b_p ($p \in \overline{0,m}$, $m \geqslant 1$). Точнее, рассмотрим гипотезу H_0^p о равенстве коэффициента b_p заданному числу b_p^0:

$$H_0^p: \; b_p = b_p^0. \tag{1.83}$$

Эта задача решается с помощью *t-статистики:*

$$t = \frac{\widehat{b}_p - b_p^0}{s_p}. \tag{1.84}$$

Если гипотеза H_0^p верна, то t-статистика имеет *распределение Стьюдента с* $(n-m-1)$ *степенями свободы.*

Задавшись уровнем значимости α, по таблицам распределения Стьюдента находят величину

$$t_{1-\alpha/2}(n-m-1), \tag{1.85}$$

т.е. квантиль распределения Стьюдента с $(n-m-1)$ степенями свободы порядка $1-\alpha/2$. Эта величина удовлетворяет условию

$$P(|t| < t_{1-\alpha/2}(n-m-1)) = 1-\alpha. \tag{1.86}$$

Тем самым определяются интервал

$$(-t_{1-\alpha/2}(n-m-1),\; t_{1-\alpha/2}(n-m-1)) \tag{1.87}$$

принятия гипотезы H_0^p и ее *критическая область*

$$\{t|\; |t| \geqslant t_{1-\alpha/2}(n-m-1)\}. \tag{1.88}$$

Если найденное по модели значение t-статистики попадает в интервал (1.87), то гипотеза H_0^p *не отвергается*; если же t – точка множества (1.88), то гипотеза *отвергается.*

Перейдем к задаче, сформулированной в заголовке этого пункта. Пусть

$$m > 1,\; p = l, l \in \overline{1,m}.$$

Теперь выбираем

$$b_p^0 = 0.$$

Поэтому t-статистика (1.84) принимает вид

$$t = \frac{\widehat{b}_l}{s_l}. \tag{1.89}$$

В соответствии с общей схемой находим наблюдаемое по модели значение статистики (1.89) и применяем разрешающее правило:

1) Если

$$|t| \geqslant t_{1-\alpha/2}(n-m-1), \tag{1.90}$$

то гипотеза

$$H_0^l:\ b_l = 0 \tag{1.91}$$

отвергается. Поэтому коэффициент $\widehat{b}_l$ является (на уровне значимости α) статистически значимым, статистически отличным от нуля. Это подтверждает целесообразность включения регрессора x_p в модель.

2) Если же

$$|t| < t_{1-\alpha/2}(n-m-1), \tag{1.92}$$

то гипотеза (1.91) не отвергается. Коэффициент $\widehat{b}_l$ считается статистически незначимым, статистически не отличающимся от нуля. Возможно, регрессор x_p следует *удалить* из модели (для принятия решения часто проводят дополнительные исследования).

1.3.5. Нахождение доверительных интервалов для коэффициентов линейной регрессии

Найденные с помощью МНК точечные оценки $\widehat{b}_p$ коэффициентов регрессии b_p[7] дополняют *интервальными оценками*. Доверительный интервал для коэффициента b_p, отвечающий доверительной вероятности (надежности) γ, запишем в виде

$$I_\gamma(b_p) = (\widehat{b}_p - \delta_p,\ \widehat{b}_p + \delta_p).$$

Чтобы найти величину δ_p, определяющую размер доверительного интервала, вновь обратимся к гипотезе (1.83). Теперь естественно истинное значение b_p^0 коэффициента b_p обозначить просто через b_p. Множество точек t искомого доверительного интервала удовлетворяет условию (1.86), где теперь

$$t = \frac{\widehat{b}_p - b_p}{s_p}.$$

[7]Теперь рассматриваются все коэффициенты регрессии, включая b_0.

Поэтому это условие запишется в виде

$$P\left(\left|\frac{\widehat{b}_p - b_p}{s_p}\right| < t_{1-\alpha/2}(n-m-1)\right) = 1-\alpha.$$

Значит, доверительный интервал $I_\gamma(b_p)$, отвечающий доверительной вероятности

$$\gamma = 1-\alpha,$$

определяется неравенством

$$\left|\frac{\widehat{b}_p - b_p}{s_p}\right| < t_{1-\alpha/2}(n-m-1).$$

Решим его относительно b_p :

$$\frac{|b_p - \widehat{b}_p|}{s_p} < t_{1-\alpha/2}(n-m-1),$$

$$\widehat{b}_p - s_p \cdot t_{1-\alpha/2}(n-m-1) < b_p < \widehat{b}_p + s_p \cdot t_{1-\alpha/2}(n-m-1).$$

Таким образом,

$$I_\gamma(b_p) = (\widehat{b}_p - \delta_p,\ \widehat{b}_p + \delta_p), \quad \delta_p = s_p \cdot t_{1-\alpha/2}(n-m-1). \tag{1.93}$$

Замечание 1.5. Доверительные интервалы для коэффициентов при регрессорах b_l, $l \in \overline{1,m}$ позволяют установить, являются ли оценки $\widehat{b}_l$ этих коэффициентов *статистически значимыми*. Гипотеза (1.91) отвергается, если

$$0 \notin I_\gamma(b_l),$$

и не отвергается, если $0 \in I_\gamma(b_l)$. Однако в эконометрической практике для решения этой задачи чаще используют подход через t-статистику, описанный в подп. 1.3.4.

Выведем расчетные формулы вычисления стандартных ошибок коэффициентов регрессии s_p для случаев $m = 1$, $m = 2$. В выкладках через $\mathbf{U}$ будем обозначать матрицу системы (1.46):

$$\mathbf{U} = \frac{1}{n}\mathbf{X}^T\mathbf{X}.$$

Пусть вначале $m = 1$. Запишем квадратную матрицу *второго* порядка $\mathbf{X}^T\mathbf{X}$, выбирая лишь две первых строки и два первых столбца матрицы общего вида (1.42) и опуская второй индекс 1:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = n \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} = n\mathbf{U}.$$

Согласно формуле (П4),

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n}\,\mathbf{U}^{-1}.$$

Для нахождения матрицы $\mathbf{U}^{-1}$ применяем формулу (П6). Так как

$$\det\,\mathbf{U} = \overline{x^2} - (\overline{x})^2 = k_{11},$$

то

$$\mathbf{U}^{-1} = \frac{1}{k_{11}} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

Поэтому

$$\mathbf{Z} = (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{nk_{11}} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

В частности,

$$z_{11} = \frac{1}{nk_{11}},\quad z_{00} = z_{11} \cdot \overline{x^2}.$$

Учитывая формулу (1.72), находим

$$s_1 = \frac{s}{\sqrt{nk_{11}}},\quad s_0 = s_1 \cdot \sqrt{\overline{x^2}} \quad (m = 1). \tag{1.94}$$

Пусть $m = 2$. Теперь

$$\mathbf{U} = \frac{1}{n}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & \overline{x}_1 & \overline{x}_2 \\ \overline{x}_1 & \overline{x_1^2} & \overline{x_1x_2} \\ \overline{x}_2 & \overline{x_1x_2} & \overline{x_2^2} \end{pmatrix}.$$

Так как

$$\mathbf{Z} = (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n}\,\mathbf{U}^{-1},$$

то

$$z_{pp} = \frac{1}{n}\,(\mathbf{U}^{-1})_{pp}.$$

Находим

$$\det \mathbf{U} \overset{(1.56)}{=} \det \mathbf{K} = \Delta = k_{11} \cdot k_{22} - k_{12}^2.$$

По формуле (П5) нахождения обратной матрицы

$$(U^{-1})_{00} = \frac{\overline{x_1^2} \cdot \overline{x_2^2} - (\overline{x_1 x_2})^2}{\Delta}, \; (U^{-1})_{11} = \frac{k_{22}}{\Delta}, \; (U^{-1})_{22} = \frac{k_{11}}{\Delta},$$

поэтому

$$z_{00} = \frac{\overline{x_1^2} \cdot \overline{x_2^2} - (\overline{x_1 x_2})^2}{n\Delta}, \; z_{11} = \frac{k_{22}}{n\Delta}, \; z_{22} = \frac{k_{11}}{n\Delta}.$$

Согласно формуле (1.72)

$$s_0 = s\sqrt{\frac{\overline{x_1^2} \cdot \overline{x_2^2} - (\overline{x_1 x_2})^2}{n\Delta}}, \; s_1 = s\sqrt{\frac{k_{22}}{n\Delta}}, \; s_2 = s\sqrt{\frac{k_{11}}{n\Delta}} \quad (m = 2). \quad (1.95)$$

1.3.6. Прогнозирование в линейных регрессионных моделях

Прогнозирование - одна из важнейших целей эконометрического моделирования.

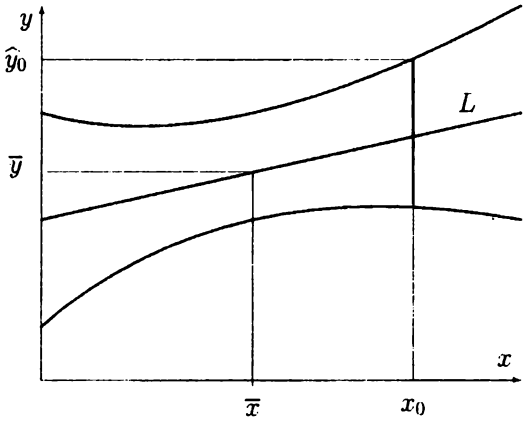


Рис. 1.3. Прогнозирование

Применительно к регрессионным моделям прогнозирование – нахождение точечной и интервальной оценок значения зависимой переменной y при *фиксированных* значениях регрессоров. Эти оценки ради

краткости будем называть точечным и интервальным прогнозом значения переменной y.

Проблематику прогнозирования обсудим на модели парной линейной регрессии (1.10).

Требуется дать прогноз $y(x_0)$ значения переменной y в точке x_0. *Точечным* прогнозом является величина

$$\widehat{y}_0 = \widehat{f}(x_0) \overset{(1.24)}{=} \overline{y} + \widehat{b}_1(x_0 - \overline{x}).$$

Геометрически точечный прогноз определяется точкой $(x_0, \widehat{y}_0)$ на прямой регрессии L (рис. 1.3).

Для нахождения интервальной оценки прогноза, т.е его доверительного интервала необходимо найти дисперсию прогноза $D(y(x_0))$. Она суммирует влияние как возможного рассеивания прямой регрессии L, так и рассеивания вокруг этой прямой.

Рассеивание прямой L определяется дисперсией случайной величины $\widehat{y}_0$. По свойствам дисперсии

$$D(\widehat{y}_0) = D(\overline{y} + \widehat{b}_1(x_0 - \overline{x})) = [\, x_0 - \overline{x} = const \,] = D(\overline{y}) + (x_0 - \overline{x})^2 D(\widehat{b}_1).$$

Вначале преобразуем

$$D(\overline{y}) = D\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2}\sum D(y_i) \overset{(1.63)}{=} \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Теперь находим

$$D(\widehat{b}_1) \overset{(1.26)}{=} D\left(\frac{k_{1y}}{k_{11}}\right) \overset{(1.55)}{=} D\left(\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}\right) =$$

$$= \frac{1}{(\sum(x_i - \overline{x})^2)^2} \cdot D\left(\sum(x_i - \overline{x})(y_i - \overline{y})\right) =$$

$$= \frac{1}{(\sum(x_i - \overline{x})^2)^2} \cdot \sum(x_i - \overline{x})^2 \cdot D(y_i - \overline{y}) \overset{(1.63)}{=}$$

$$= \frac{1}{(\sum(x_i - \overline{x})^2)^2} \cdot \sigma^2 \sum(x_i - \overline{x})^2 = \frac{\sigma^2}{\sum(x_i - \overline{x})^2} \overset{(1.55)}{=} \frac{\sigma^2}{nk_{11}}.$$

Следовательно,

$$D(\widehat{y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nk_{11}}\right).$$

Учитывая, что рассеивание вокруг прямой L привносит в дисперсию слагаемое σ^2, можем записать

$$D(y(x_0)) - \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nk_{11}}\right).$$

Заменяя, как и ранее, σ на s и извлекая квадратный корень, приходим к *стандартной ошибке прогноза*

$$s_{\text{пр}} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nk_{11}}}. \tag{1.96}$$

Окончательно доверительный интервал прогноза задается так:

$$I_\gamma(y(x_0)) = (\widehat{y}_0 - \delta_{\text{пр}},\ \widehat{y}_0 + \delta_{\text{пр}}),\ \delta_{\text{пр}} = s_{\text{пр}} \cdot t_{1-\alpha/2}(n-2). \tag{1.97}$$

Размер интервала, помимо стандартной ошибки прогноза $s_{\text{пр}}$, определяется также квантилем $t_{1-\alpha/2}(n-2)$ распределения Стьюдента с $(n-2)$ степенями свободы (здесь $m=1$, поэтому $n-m-1=n-2$).

Из формулы (1.96) видно, что стандартная ошибка прогноза $s_{\text{пр}}$ *минимальна* при $x_0 = \overline{x}$ и *возрастает* при удалении x_0 от $\overline{x}$ (как влево, так и вправо). Поэтому при таком удалении ширина доверительного интервала увеличивается, а точность прогноза уменьшается.

1.3.7. Примеры статистического исследования регрессионных моделей

Проведем статистическое исследование линейных регрессионных моделей примера 1.1. (парная регрессия, $n=8$, $m=1$) и примера 1.2. (трехмерная регрессия, $n=10$, $m=2$). Выбираем уровень значимости $\alpha = 0,05$, тогда $\gamma = 1-\alpha = 0,95$.

Пример 1.1. (продолжение)

Требуется:

3) С помощью коэффициента детерминации и F-статистики оценить качество модели в целом.

Если модель в целом соответствует наблюдениям, то следует:

4) Вычислить стандартную ошибку регрессии и стандартные ошибки коэффициентов регрессии.

5) Найти доверительные интервалы для коэффициентов регрессии.

6) Найти точечный и интервальный прогнозы $y(140)$ для нового магазина с персоналом численностью $x_0 = 140$ (чел.).

Решение:

3) Найдем коэффициент детерминации

$$R^2 \overset{(1.81)}{=} \frac{k_{1y}^2}{k_{11}k_{yy}} = \frac{10,475^2}{544,5 \cdot 0,2075} = 0,971.$$

Вычисляем

$$F_{\text{набл}} \overset{(1.82)}{=} [m = 1] = \frac{(n-2)R^2}{1-R^2} = \frac{6 \cdot 0,971}{0,029} = \frac{5,8272}{0,029} = 202,33.$$

По таблицам распределения Фишера находим

$$F_{\text{крит}} = F_{1-\alpha}(m, n-m-1) = F_{0,95}(1,6) = 5,99 < F_{\text{набл}}.$$

Так как $F_{\text{крит}} < F_{\text{набл}}$, то построенная линейная модель соответствует наблюдениям (на уровне значимости $\alpha = 0,05$).

4) Последовательно вычисляем:

$$s = \sqrt{s^2} \overset{(1.81)}{=} \sqrt{\frac{n(k_{yy} - \widehat{b}_1 k_{1y})}{n-2}} = \sqrt{\frac{8 \cdot (0,2075 - 0,0192 \cdot 10,475)}{6}} =$$

$$= \sqrt{\frac{4}{3} \cdot (0,2075 - 0,2011)} = 0,09;$$

$$s_1 \overset{(1.94)}{=} \frac{s}{\sqrt{nk_{11}}} = \frac{0,09}{\sqrt{8 \cdot 544,5}} = \frac{0,09}{66} = 0,0014;$$

$$s_0 \overset{(1.94)}{=} s_1 \cdot \sqrt{\overline{x^2}} = 0,0014 \cdot \sqrt{13313,5} = 0,0014 \cdot 115,38414 = 0,162.$$

5) По таблицам распределения Стьюдента

$$t_{1-\alpha/2}(n-2) = t_{0,975}(6) = 2,45.$$

Находим

$$\delta_1 = s_1 \cdot t_{1-\alpha/2}(n-2) = 0,0014 \cdot 2,45 = 0,0034;$$

$$\delta_0 = s_0 \cdot t_{1-\alpha/2}(n-2) = 0,162 \cdot 2,45 = 0,397.$$

Поэтому доверительные интервалы для параметров b_1, b_0 имеют вид

$$I_{0,95}(b_1) = (\widehat{b}_1 - \delta_1,\ \widehat{b}_1 + \delta_1) =$$

$$= (0,0192 - 0,0034;\ 0,0192 + 0,0034) = (0,0158; 0,0226);$$

$$I_{0,95}(b_0) = (\widehat{b}_0 - \delta_0,\ \widehat{b}_0 + \delta_0) =$$

$$= (-0,973 - 0,397;\ -0,973 + 0,397) = (-1,37;\ -0,576).$$

6) Найдем точечную оценку прогноза (см. рис. 1.2):

$$\widehat{y}_0 = \widehat{y}_{140} = \widehat{b}_0 + \widehat{b}_1 \cdot 140 = -0,973 + 0,0192 \cdot 140 = -0,973 + 2,688 = 1,715.$$

Вычислим теперь стандартную ошибку прогноза:

$$s_{\text{пр}} \overset{(1.96)}{=} s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nk_{11}}} = 0,09 \cdot \sqrt{1 + \frac{1}{8} + \frac{(140 - 113)^2}{8 \cdot 544,5}} =$$

$$= 0,09 \cdot \sqrt{1,125 + \frac{729}{4356}} = 0,09 \cdot \sqrt{1,125 + 0,167} =$$

$$= 0,09 \cdot 1,137 = 0,102$$

и величину $\delta_{\text{пр}}$:

$$\delta_{\text{пр}} = s_{\text{пр}} \cdot t_{1-\alpha/2}(n-2) = 0,102 \cdot 2,45 = 0,25.$$

Поэтому доверительный интервал прогноза

$$I_{0,95}(y(140)) \overset{(1.97)}{=} (\widehat{y}_0 - \delta_{\text{пр}},\ \widehat{y}_0 + \delta_{\text{пр}}) =$$

$$= (1,715 - 0,25;\ 1,715 + 0,25) = (1,465;\ 1,965).$$

Таким образом, линейная модель прогнозирует с надежностью $\gamma = 0,95$, что магазин с числом сотрудников 140 чел. может иметь товарооборот в диапазоне от 1,465 до 1,965 млн р.

Пример 1.2. (продолжение)

Требуется:

2) С помощью коэффициента детерминации и F-статистики оценить качество модели в целом.

Если модель в целом соответствует наблюдениям, то следует:

3) Вычислить стандартные ошибки коэффициентов регрессии.

4) Оценить статистическую значимость коэффициентов $\widehat{b}_1$, $\widehat{b}_2$ при регрессорах.

5) Найти доверительные интервалы для всех коэффициентов регрессии b_0, b_1, b_2.

Решение:

2) Для проверки соответствия модели результатам наблюдений найдем коэффициент детерминации

$$R^2 \overset{(1.78)}{=} \frac{\widehat{b}_1 k_{1y} + \widehat{b}_2 k_{2y}}{k_{yy}} = \frac{0,117 \cdot 375 + 3,447 \cdot 7,7}{72,4} = \frac{70,417}{72,4} = 0,973.$$

Вычисляем

$$F_{\text{набл}} \overset{(1.82)}{=} \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = [\, n = 10, m = 2 \,] = \frac{0,973 \cdot 7}{0,027 \cdot 2} = 126,13.$$

По таблицам распределения Фишера находим

$$F_{\text{крит}} = F_\alpha(m, n-m-1) = F_{0,95}(2,7) = 4,74 < F_{\text{набл}}.$$

Следовательно, построенная линейная модель в целом соответствует наблюдениям.

3) Применяя формулы (1.95), (1.80), вычисляем стандартные ошибки коэффициентов регрессии:

$$s_0 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot (\overline{x_1^2} \cdot \overline{x_2^2} - (\overline{x_1 x_2})^2)} =$$

$$= \sqrt{\frac{72,4 \cdot 0,027}{7 \cdot 613,4} \cdot (30320 \cdot 11,2 - 575^2)} = \sqrt{\frac{17513,05}{4293,8}} = 2,02;$$

$$s_1 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot k_{22}} =$$

$$= \sqrt{\frac{72,4 \cdot 0,027}{7 \cdot 613,4} \cdot 0,96} = \sqrt{\frac{1,9548}{4293,8}} = 0,021;$$

$$s_2 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot k_{11}} = \sqrt{\frac{72,4 \cdot 0,027}{7 \cdot 613,4} \cdot 2096} = \sqrt{\frac{4097,26}{4293,8}} =$$

$$= \sqrt{0,9542} = 0,977.$$

4) По таблицам квантилей распределения Стьюдента находим

$$t_{1-\alpha/2}(n-m-1) = t_{0,975}(7) = 2,36.$$

Для оценивания статистической значимости коэффициента $\widehat{b}_1$ сравним наблюдаемое по модели значение статистики (1.89) со значением $t_{0,975}(7) = 2,36$:

$$t = \frac{\widehat{b}_1}{s_1} = \frac{0,117}{0,021} = 5,57 > 2,36.$$

Аналогично работаем с коэффициентом при другом регрессоре:

$$t = \frac{\widehat{b}_2}{s_2} = \frac{3,447}{0,977} = 3,53 > 2,36.$$

Так как в обоих случаях

$$|t| \geqslant t_{0,975}(7),$$

то коэффициенты $\widehat{b}_1, \widehat{b}_2$ статистически значимы. Поэтому включение в модель переменных x_1, x_2 оправданно.

5) Находим

$$\delta_0 \overset{(1.93)}{=} s_0 \cdot t_{0,975}(7) = 2,02 \cdot 2,36 = 4,767;$$

$$\delta_1 = s_1 \cdot t_{0,975}(7) = 0,021 \cdot 2,36 = 0,05;$$

$$\delta_2 = s_2 \cdot t_{0,975}(7) = 0,977 \cdot 2,36 = 2,306.$$

Поэтому

$$I_{0,95}(b_0) = (\widehat{b}_0 - \delta_0,\ \widehat{b}_0 + \delta_0) =$$

$$= (4,314 - 4,767;\ 4,314 + 4,767) = (-0,453; 9,081);$$

$$I_{0,95}(b_1) = (0,117 - 0,05;\ 0,117 + 0,05) = (0,067; 0,167);$$

$$I_{0,95}(b_2) = (3,447 - 2,306;\ 3,447 + 2,306) = (1,141;\ 5,753).$$

1.4. Фиктивные переменные

Мы рассматривали линейные регрессионные модели, где регрессоры являлись *количественными* переменными, значения которых – действительные числа. Однако на практике часто приходится учитывать влияние и *классификационных* факторов [21, с. 5][8]. Примеры таких

[8]В эконометрической литературе их часто называют *качественными* факторами.

факторов — пол человека, уровень его образования, время года, день недели. Классификационные факторы (переменные) позволяют разбивать наблюдаемые объекты на конечное число непересекающихся *классов*. Так, классификационный признак «пол человека» делит рассматриваемое множество людей на два класса – мужчин и женщин, он имеет две градации.

Введение классификационных признаков может существенно влиять на структуру связей между изучаемыми переменными и приводить к скачкообразному изменению их характеристик. В таких случаях говорят о регрессионных моделях *с переменной структурой*.

Пусть, например, требуется изучить зависимость заработной платы y работников фирмы от количественных факторов $x_1, \ldots, x_m$ и от классификационного фактора z (пол работника). Если не рассматривать фактор z, то приходим к обычной модели линейной регрессии

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_m x_{im} + \varepsilon_i \quad (i \in \overline{1,n}). \tag{1.98}$$

Учесть фактор z можно следующим образом: расщепив модель (1.98) на две подмодели — для работников-мужчин и работников-женщин, исследовать эти подмодели отдельно, а затем выявить различия между ними.

В эконометрике принят другой подход, позволяющий изучать влияние всех факторов – количественных и классификационных – с помощью *одного* уравнения регрессии. Эффект классификационных факторов при этом учитывается введением фиктивных переменных (манекенов). Принято в качестве фиктивных переменных выбирать бинарные (булевы) переменные, принимающие лишь два значения – 0 и 1.

Если классификационный признак имеет две градации, то вводят одну фиктивную переменную. Если же классификационный признак принимает k ($k > 2$) значений, то задают $(k-1)$ фиктивных переменных.

Вернемся к нашему примеру. Введем фиктивную переменную z, учитывающую влияние фактора «пол работника», например, так:

$$z_i = \begin{cases} 1, & \text{если } i\text{-й работник – мужчина,} \\ 0, & \text{если } i\text{-й работник – женщина.} \end{cases} \tag{1.99}$$

Тогда модель (1.98) преобразуется к виду

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_m x_{im} + c z_i + \varepsilon_i \quad (i \in \overline{1,n}). \tag{1.100}$$

Она содержит $(m+1)$ регрессоров. Новый коэффициент c интерпретируется следующим образом. Средняя зарплата у мужчин отличается от средней зарплаты у женщин на c единиц (при неизменных значениях остальных регрессоров). Проверяя гипотезу

$$H_0 : \ c = 0,$$

мы можем установить статистическую существенность или статистическую несущественность влияния фактора z на размер зарплаты работника фирмы. Если гипотеза H_0 отклоняется, то на фирме имеется дискриминация в оплате по половому признаку: при $c > 0$ – в пользу мужчин, при $c < 0$ – в пользу женщин.

Пример 1.3. На предприятии используются станки двух фирм – A и B. Исследуется надежность станков, т.е. зависимость времени безаварийной работы y (в часах) от возраста станка x (в месяцах). Результаты выборки для 16 станков приведены в табл. 1.5[9].

Таблица 1.5

Фирма	A	B	A	B	A	B	A	B
x_{1i}	23	65	69	75	63	75	52	70
y_i	280	112	176	90	176	110	200	148
Фирма	B	A	A	B	A	A	A	A
x_{1i}	62	66	20	39	48	59	25	71
y_i	150	123	245	176	236	205	240	115

Требуется:

1) На плоскости (x, y) построить корреляционное поле, отмечая различными символами точки поля, соответствующие станкам разных фирм.

2) Рассмотреть линейную регрессионную модель зависимости y от x с учетом фактора z «фирма – изготовитель станка». Найти точечные оценки коэффициентов модели. Построить на плоскости (x, y) частные прямые регрессий y по x: для станков фирмы A и для станков фирмы B в отдельности.

[9]См. работу С.А. Бородича [3, с. 304].

3) С помощью коэффициента детерминации R^2 на уровне значимости $\alpha = 0,05$ проверить соответствие регрессионной модели наблюдениям в целом.

4) В случае соответствия установить, является ли фактор z существенным для переменной y.

Решение:

1) Имеем 10 наблюдений над станками фирмы A и 6 наблюдений над станками фирмы B. Построим на плоскости (x, y) корреляционное поле. Точки поля, отвечающие станкам фирмы A, отмечаем кружочками, а точки, соответствующие станкам фирмы B — зачерненными кружочками (рис. 1.4). Видно, что точки, отмеченные зачерненными кружочками, в среднем расположены ниже «светлых» кружочков; возможно, надежность станков фирмы B меньше. В дальнейшем мы это предположение обоснуем.

2) Построим линейную регрессионную модель с зависимой переменной y и двумя регрессорами $x_1 = x$, $x_2 = z$:

$$y = b_0 + b_1 x + cz + \varepsilon. \tag{1.101}$$

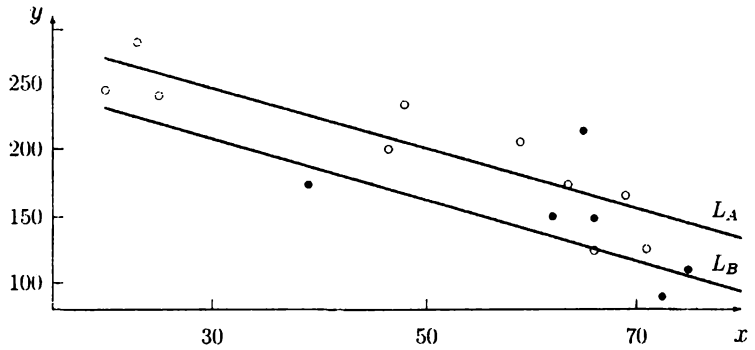


Рис. 1.4. Фиктивные переменные. Пример 1.3

Регрессор z, являющийся фиктивной переменной, определим следующим образом:

$$z_i = \begin{cases} 0, \text{ если } i\text{-й станок произведен на фирме } A, \\ 1, \text{ если } i\text{-й станок произведен на фирме } B. \end{cases} \tag{1.102}$$

Запишем нашу модель в матричной форме (1.30). Здесь

$$\mathbf{y} = \mathbf{y}_{16\times 1} = \begin{pmatrix} 280 \\ 112 \\ \dots \\ 240 \\ 115 \end{pmatrix}, \mathbf{X} = \mathbf{X}_{16\times 3} = \begin{pmatrix} 1 & 23 & 0 \\ 1 & 65 & 1 \\ 1 & 69 & 0 \\ 1 & 75 & 1 \\ \dots & \dots & \dots \\ 1 & 48 & 0 \\ 1 & 59 & 0 \\ 1 & 25 & 0 \\ 1 & 71 & 0 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ c \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{16\times 1} = (\varepsilon_1, \dots, \varepsilon_{16})^T.$$

Укажем значения параметров модели, необходимых для нахождения коэффициентов регрессии[10]:

$$\overline{x}_1 = 55,125;\ \overline{x_1^2} = 3369,375;\ \overline{x}_2 = \overline{x_2^2} = 0,375;\ \overline{y} = 173,875;\ \overline{y^2} = 33261;$$

$$\overline{x_1 x_2} = 24,125;\ \overline{x_1 y} = 8717,625;\ \overline{x_2 y} = 49,125;$$

$$k_{11} = 330,6094;\ k_{12} = 3,4531;\ k_{22} = 0,2344;\ k_{1y} = -867,2344;$$

$$k_{2y} = -16,0781;\ k_{yy} = 3028,4844;\ \Delta = 65,5625.$$

По формулам (1.54), где $\widehat{b}_2$ заменяется на $\widehat{c}$, находим точечные оценки коэффициентов регрессии:

$$\widehat{b}_0 = 311,368;\ \widehat{b}_1 = -2,253;\ \widehat{c} = -35,4.$$

Поэтому уравнение линейной регрессии y на x, z записывается следующим образом:

$$\widehat{y} = 311,368 - 2,253x - 35,4z. \tag{1.103}$$

Укажем экономический смысл коэффициентов $\widehat{b}_1$, $\widehat{c}$. При увеличении возраста станка на 1 месяц время его безаварийной работы уменьшается в среднем на 2,553 часа (для всех станков предприятия). Так как $\widehat{c} = -35,4$, то время безаварийной работы станков фирмы B меньше времени безаварийной работы станков фирмы A в среднем на 35,4 часа.

[10]Параметры вычислены с использованием математического пакета MathCAD.

Найдем *частные уравнения регрессии* для станков фирм A, B, полагая в уравнении (1.103) $z = 0, z = 1$ соответственно:

$$\widehat{y} = 311,368 - 2,253x;$$

$$\widehat{y} = 275,968 - 2,253x.$$

Эти уравнения определяют на плоскости (x, y) параллельные прямые

$$L_A: \quad y = 311,368 - 2,253x,$$

$$L_B: \quad y = 275,968 - 2,253x.$$

Прямая L_B расположена ниже прямой L_A на $|\widehat{c}| = 35,4$ единицы.

3) Найдем коэффициент детерминации

$$R^2 \overset{(1.81)}{=} 0,8332.$$

Вычисляем

$$F_{\text{набл}} \overset{(1.82)}{=} 32,469.$$

По таблицам распределения Фишера определяем

$$F_{\text{крит}} = F_\alpha(m, n-m-1) = F_{0,95}(2,13) = 3,81 < F_{\text{набл}}.$$

Поэтому модель (1.101) в целом соответствует наблюдениям.

4) Проверим, является ли классификационный фактор z статистически существенным для переменной y. Иначе говоря, выясним, являются ли частные прямые регрессии L_A, L_B статистически различными. С этой целью проверяем гипотезу $H_0: c = 0$. Она отвергается, если

$$|t| = \left|\frac{\widehat{c}}{s_2}\right| \geqslant t_{1-\alpha/2}(n-m-1). \tag{1.104}$$

Находим последовательно

$$s_2 \overset{(1.95)}{=} 13,997;\ t_{1-\alpha/2}(n-m-1) = t_{0,975}(n-3) = t_{0,975}(13) = 2,16.$$

Поэтому

$$|t| = \left|\frac{-35,4}{13,997}\right| = 2,529 > 2,16$$

и условие (1.104) выполнено. Значит, фактор z является статистически значимым для переменной y.

1.5. Неклассические случаи линейной регрессии

Применяя при анализе линейных регрессионных моделей МНК, необходимо выяснить, выполнены или нет условия Гаусса–Маркова. В случае их нарушения этот метод может давать оценки с плохими статистическими свойствами. Сейчас рассмотрим способы устранения или смягчения нарушений некоторых условий Гаусса–Маркова, а также тесты, позволяющие выявлять эти нарушения.

1.5.1. Обобщенная линейная регрессионная модель, обобщенный метод наименьших квадратов

На практике встречаются ситуации, когда не выполняется условие (1.61) классической линейной регрессионной модели. Напомним, что оно означает постоянство дисперсий (гомоскедастичность) случайных ошибок регрессии ε_i и их некоррелированность. Эти же свойства справедливы и для наблюдаемых значений y_i зависимой переменной y.

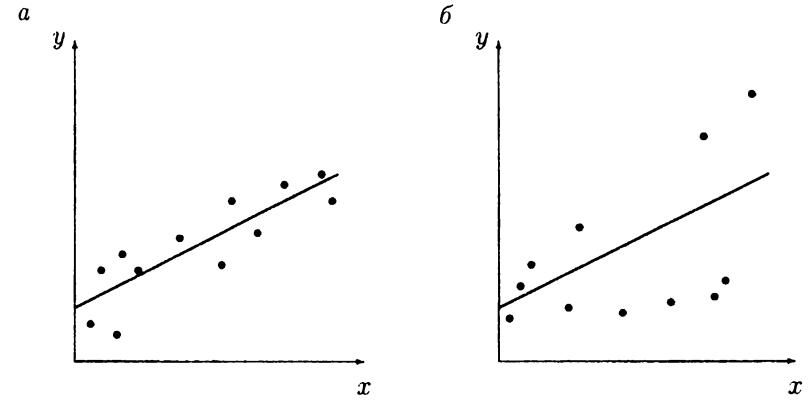


Рис. 1.5. Различный характер корреляционного поля:
а – гомоскедастичность, *б* – гетероскедастичность

Требование гомоскедастичности оправданно, когда наблюдения y_i достаточно однородны. Однако в ряде ситуаций данное предположение нереалистично. Так, если исследуется зависимость расходов на питание

в семьях от их общего дохода, то естественно считать, что разброс в семьях с более высокими доходами больше. Значит, дисперсии случайных величин y_i непостоянны и мы приходим к *гетероскедастичности* (так называют нарушение свойства гомоскедастичности) (рис.1.5).

При анализе временных рядов наблюдения y_i за переменной y зачастую коррелируют с предыдущими наблюдениями (y_{i-1}, y_{i-2} и т.д.), что и означает автокорреляцию.

Рассмотрим теперь *обобщенную линейную регрессионную модель*, отличающуюся от классической линейной модели (см. подп.1.3.1) лишь тем, что вместо условия (1.61) вводится более общее условие

$$\mathbf{K}(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega}. \tag{1.105}$$

Здесь $\boldsymbol{\Omega} = (\omega_{ij})$ заданная положительно определенная (симметрическая) матрица порядка n.

Для обобщенной модели, как и для классической, можно рассматривать точечную МНК-оценку (1.37) вектора $\mathbf{b}$. Она и теперь не смещена и состоятельна. Но для обобщенной модели оценка (1.37) неэффективна, при этом важная для качества оценивания матрица $\mathbf{K}(\widehat{\mathbf{b}})$ определяется по формуле

$$\mathbf{K}(\widehat{\mathbf{b}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1},$$

существенно отличной от формулы (1.67). Оценки дисперсий коэффициентов $\widehat{b}_i$ будут смещенными, что может привести к неверным выводам о качестве регрессии.

Оказывается, эффективной является другая точечная оценка вектора $\mathbf{b}$, зависящая от матрицы $\boldsymbol{\Omega}$:

$$\widehat{\mathbf{b}}_0 = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{y}. \tag{1.106}$$

Ее называют точечной оценкой *обобщенного метода наименьших квадратов* (ОМНК). Суть ОМНК состоит в решении *обобщенной* экстремальной задачи

$$Q_0(\mathbf{b}) = \mathbf{e}^T\boldsymbol{\Omega}^{-1}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}) \to \min \tag{1.107}$$

[12, с. 152--155].

Обобщенная линейная регрессионная модель имеет ту особенность, что коэффициент детерминации R^2 не может служить показателем качества модели. Значение коэффициента даже может выйти за пределы отрезка $[0.1]$.

Параметрами обобщенной модели являются коэффициенты регрессии $b_0, b_1, \ldots, b_m$ и элементы матрицы $\mathbf{\Omega}$. Поскольку эта матрица симметрическая и $\omega_{ij} = \omega_{ji}$, то можем рассматривать лишь $n(n+1)/2$ ее элементов, расположенных не ниже главной диагонали. Таким образом, число всех параметров модели – $(n(n+1)/2 + m + 1)$. Так как $n(n+1)/2 + m + 1 > n$, то оценить все параметры по n наблюдениям невозможно. Поэтому в общей постановке ОМНК нереализуем. Для практического применения метода ОМНК необходимо вводить *априорные ограничения на структуру матрицы* $\mathbf{\Omega}$, причем желательно, чтобы она содержала небольшое число оцениваемых параметров.

Рассмотрим два реалистичных варианта ОМНК, отвечающие гетероскедастичности и автокорреляции ошибок регрессии.

1.5.2. Гетероскедастичность, взвешенный метод наименьших квадратов

Существует несколько способов обнаружения гетероскедастичности, использующих различные тесты – процедуры статистической проверки гипотез. Обычно в тестах в качестве основной проверяется гипотеза H_0 *об отсутствии гетероскедастичности.* Многие тесты предполагают определенные ограничения на характер гетероскедастичности. Рассмотрим применительно к парной регрессии часто используемый *тест Голдфелда - Квандта.* Будем считать, что ошибки регрессии ε_i нормально распределены, а их дисперсии $D(\varepsilon_i)$ - возрастающие функции регрессора[11]. Тест опирается на сравнение дисперсий двух подвыборок и включает следующие этапы:

1) Все n наблюдений упорядочиваются по величине регрессора x.

2) Полученная упорядоченная выборка разбивается (без изменения порядка) на три подвыборки размерностей $k, n-2k, k$ соответственно $(k \approx n/3, k > n/3)$.

3) Классическим методом наименьших квадратов исследуются линейные регрессии, отвечающие двум крайним подвыборкам, и находятся суммы квадратов

$$S_1 = \sum_{i=1}^{k} e_i^2, \; S_2 = \sum_{i=n-k+1}^{n} e_i^2.$$

[11]Часто предполагают, что эта зависимость квадратична: $D(\varepsilon_i) = c^2 x_i^2$.

Если гетероскедастичность предполагаемого характера имеется, то S_2 существенно больше S_1.

4) Рассматривается статистика

$$F = \frac{S_2}{S_1}.$$

В случае истинности гипотезы H_0 она имеет распределение Фишера со степенями свободы $k_1 = k_2 = k - m - 1$ (здесь m – число регрессоров, в рассматриваемом случае $m = 1$).

5) Если

$$\frac{S_2}{S_1} > F_{1-\alpha}(k_1, k_2)\,,$$

то гипотеза H_0 отклоняется в пользу гетероскедастичности.

Тест можно использовать и в предположении об обратной пропорциональности между $D(\varepsilon_i), x_i^2$, тогда

$$F = \frac{S_1}{S_2}.$$

В случае гетероскедастичности дисперсии ошибок $D(\varepsilon_i)$ различны:

$$D(\varepsilon_i) = \sigma_i^2 \quad (i \in \overline{1,n}).$$

Так как 4^0 условие Гаусса-Маркова предполагается выполненным, то матрица $\mathbf{\Omega}$ диагональна:

$$\mathbf{\Omega} = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\} = \begin{pmatrix} \sigma_1^2 & 0 & & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}. \tag{1.108}$$

Тогда обратная матрица также диагональна и легко находится:

$$\mathbf{\Omega}^{-1} = diag\left\{\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right\}.$$

Поэтому применение ОМНК в рассматриваемом случае означает решение экстремальной задачи

$$Q_0(\mathbf{b}) = \mathbf{e}^T\mathbf{\Omega}^{-1}\mathbf{e} = \sum_{i=1}^{n}\left(\frac{e_i}{\sigma_i}\right)^2 \to \min\,.$$

Здесь минимизируется *взвешенная* сумма квадратов: остатки регрессии e_i входят с множителями – весами $1/\sigma_i$. Тем самым менее точным наблюдениям, имеющим большие дисперсии ошибок, придается малый вес и, наоборот, наблюдениям с малыми дисперсиями ошибок – большой вес. Модификация ОМНК применительно к случаю (1.108) называется *взвешенным методом наименьших квадратов.*

Этот вариант обобщенной линейной регрессионной модели содержит $(m + n + 1)$ параметров

$$b_0, b_1, \ldots, b_m, \sigma_1^2, \ldots, \sigma_n^2.$$

1.5.3. Автокорреляция

Автокорреляция – зависимость (коррелированность) в регрессионном уравнении между наблюдаемыми значениями y_i зависимой переменной y или, что то же самое, между ошибками регрессии ε_i. Поэтому автокорреляция означает невыполнение 4^0 условия Гаусса–Маркова.

Последствия автокорреляции в определенной степени схожи с последствиями гетероскедастичности: оценки коэффициентов регрессии неэффективны, их стандартные ошибки оцениваются неправильно (во многих случаях занижаются). Последнее приводит к излишне оптимистическим выводам о модели.

Рассмотрим (для простоты) парную линейную регрессию (1.10)

$$y_i = b_0 + b_1 x_i + \varepsilon_i \ (i \in \overline{1,n}). \tag{1.109}$$

Чаще всего автокорреляция возникает в регрессионных моделях с временными выборками, т.е. во временных рядах. Если же модель (1.109) соответствует пространственной выборке, то выборку необходимо упорядочить по возрастанию регрессора.

В случае автокорреляции наибольшее влияние на наблюдение y_i за переменной y обычно оказывает *предыдущее* наблюдение y_{i-1}. Эффект от более ранних наблюдений (y_{i-2} и т.д.) часто не учитывают. Такую автокорреляцию называют *автокорреляцией первого порядка.* Ее удобно формализовать как следующую зависимость между соседними ошибками регрессии $\varepsilon_i, \varepsilon_{i-1}$:

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i \quad (|\rho| < 1). \tag{1.110}$$

Число ρ называется *коэффициентом автокорреляции*, случайные ошибки u_i удовлетворяют условиям Гаусса–Маркова.

Для выявления автокорреляции первого порядка часто применяют тест Дарбина–Уотсона (Durbin–Watson). Он основан на следующей идее: если корреляция ошибок регрессии существует, то она присутствует и в остатках регрессии e_i, получающихся с применением обычного МНК. В тесте используется статистика Дарбина–Уотсона:

$$DW = \frac{\sum\limits_{i=2}^{n}(e_i - e_{i-1})^2}{\sum\limits_{i=1}^{n} e_i^2}. \tag{1.111}$$

Будем предполагать, что размерность n выборки достаточно велика. Убедимся, что статистика DW связана с выборочным коэффициентом r корреляции между соседними наблюдениями. Вначале преобразуем

$$r = \frac{\sum\limits_{i=2}^{n}(e_i - M(e_i))(e_{i-1} - M(e_{i-1}))}{\sqrt{\sum\limits_{i=1}^{n} e_i^2 \cdot \sum\limits_{i=2}^{n} e_{i-1}^2}} =$$

$$= \left[M(e_i) = M(e_{i-1}) = 0, \sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2 \right] \approx \frac{\sum\limits_{i=2}^{n} e_i e_{i-1}}{\sum\limits_{i=1}^{n} e_i^2}.$$

Отсюда

$$DW = \frac{\sum\limits_{i=2}^{n} e_i^2 + \sum\limits_{i=2}^{n} e_{i-1}^2 - 2\sum\limits_{i=2}^{n} e_i e_{i-1}}{\sum\limits_{i=1}^{n} e_i^2} =$$

$$= \frac{\sum\limits_{i=1}^{n} e_i^2 - e_1^2 + \sum\limits_{i=1}^{n} e_i^2 - e_n^2 - 2\sum\limits_{i=2}^{n} e_i e_{i-1}}{\sum\limits_{i=1}^{n} e_i^2} = 2\left(1 - \frac{\sum\limits_{i=2}^{n} e_i e_{i-1}}{\sum\limits_{i=1}^{n} e_i^2}\right) -$$

$$- \frac{e_1^2 + e_n^2}{\sum\limits_{i=1}^{n} e_i^2} \approx \left[\frac{e_1^2 + e_n^2}{\sum\limits_{i=1}^{n} e_i^2} \approx 0 \right] \approx 2(1 - r).$$

Таким образом, получили приближенную простую формулу

$$DW \approx 2(1-r). \tag{1.112}$$

Поскольку $-1 \leqslant r \leqslant 1$, то $0 \leqslant 1-r \leqslant 2$ и $DW \in [0,4]$. В случае отсутствия автокорреляции коэффициент r близок к нулю, а статистика DW близка к двум. В тесте Дарбина–Уотсона используются нижняя граничная точка d_1 и верхняя граничная точка d_2 статистики DW:

$$d_1 = d_1(\alpha, n, m),\ d_2 = d_2(\alpha, n, m) \quad (0 < d_1 < d_2 < 2).$$

Рассматривается основная гипотеза H_0 : *автокорреляция отсутствует.*

Вычисляется наблюдаемое значение $DW_{\text{набл}} = d$ статистики Дарбина–Уотсона. В зависимости от значения d делаются следующие выводы:

1) Если $d \in [0, d_1) \cup (4-d_1, 4]$, то гипотеза H_0 на уровне значимости α отвергается, автокорреляция существует.

2) Если $d \in [d_1, d_2] \cup [4-d_2, 4-d_1]$, то тест не работает.

3) Если $d \in (d_2, 4-d_2)$, то гипотеза об отсутствии автокорреляции принимается.

Автокорреляция первого порядка устраняется *авторегрессионным преобразованием*[12]. Для его вывода преобразуем разность

$$y_i - \rho y_{i-1} \overset{(1.109)}{=} b_0 + b_1 x_i + \varepsilon_i - \rho(b_0 + b_1 x_{i-1} + \varepsilon_{i-1}) =$$
$$= b_0(1-\rho) + b_1(x_i - \rho x_{i-1}) + (\varepsilon_i - \rho\varepsilon_{i-1}) = b_0(1-\rho) + b_1(x_i - \rho x_{i-1}) + u_i.$$

Произведя замены

$$y_i' = y_i - \rho y_{i-1},\ x_i' = x_i - \rho x_{i-1}, \tag{1.113}$$

$$b_0' = b_0(1-\rho), \tag{1.114}$$

приходим к *классической* линейной регрессионной модели

$$y_i' = b_0' + b_1 x_i' + u_i \quad (i \in \overline{2,n}). \tag{1.115}$$

Оценив ее, сможем по формуле (1.114) пересчитать оценку коэффициента b_0 исходной модели (1.109).

---

[12]Происхождение термина объясняется тем, что в теории временных рядов модель (1.110) называется авторегрессионной (см. подп. 3.2.1).

Таким образом, при известном значении коэффициента ρ автокорреляция устраняется. На практике, однако, коэффициент ρ неизвестен, поэтому находят его оценку. Наиболее простой оценкой, пригодной при достаточно больших выборках, является приближенная формула

$$\rho \approx r.$$

Замечание 1.6. Если рассматривается модель линейной *множественной* регрессии (1.30) с автокорреляцией первого порядка (1.110), то и к ней может быть применено авторегрессионное преобразование, аналогичное (1.113). Такой подход равносилен реализации в модели (1.30) ОМНК с матрицей

$$\mathbf{\Omega} = \frac{\sigma_u^2}{1-\rho}\begin{pmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{pmatrix}.$$

Здесь $\sigma_u^2 = D(u_i)$.

Пример 1.4[13]. Известна динамика изменения в течение 16 лет среднедушевых расходов населения некоторой страны на потребление y_i в зависимости от среднедушевого дохода x_i (табл. 1.6).

Таблица 1.6

i	1	2	3	4	5	6	7	8
x_i	73	76	83	89	95	100	107	108
y_i	70	73	78	83	86	89	96	96
i	9	10	11	12	13	14	15	16
x_i	113	119	121	122	131	135	139	140
y_i	103	109	112	114	115	118	122	123

Требуется:

1) Найти точечные оценки коэффициентов $\widehat{b}_0, \widehat{b}_1$ линейной регрессии (1.109). Используя остатки регрессии, применить тест Дарбина–Уотсона на уровне значимости $\alpha = 0,05$.

[13] См. работу А.И. Новикова [16, с. 77].

2) В случае принятия гипотезы об автокорреляции уточнить исходную модель с помощью автокорреляционного преобразования. Применить тест Дарбина–Уотсона к преобразованной модели. Если автокорреляция в ней отсутствует, то пересчитать коэффициент $\widehat{b}_0$.

Решение:

1) Используя выборку, находим точечные оценки коэффициентов парной линейной регрессии (1.109)

$$\widehat{b}_0 = 10,987;\ \widehat{b}_1 = 0,806.$$

Поэтому оцененная функция регрессии записывается в виде

$$\widehat{y} = 10,987 + 0,806\,x.$$

Таблица 1.7

i	x_i	y_i	e_i	x'_i	y'_i	e'_i
1	73	70	0,179	–	–	–
2	76	73	0,761	38,989	37,51	0,437
3	83	78	0,12	44,468	40,989	– 0,419
4	89	83	0,283	46,919	43,454	0,107
5	95	86	– 1,552	49,877	43,919	– 1,768
6	100	89	– 2,581	51,835	45,398	– 1,838
7	107	96	1,223	56,3	50,877	0,109
8	108	96	– 2,029	53,751	47,328	1,424
9	113	103	0,941	58,244	54,328	2,022
10	119	109	2,106	61,709	56,779	1,732
11	121	112	3,494	60,667	56,737	2,514
12	122	114	4,688	60,653	57,216	3,004
13	131	115	– 1,566	69,146	57,202	– 3,729
14	135	118	– 1,789	68,583	59,695	– 0,79
15	139	122	– 1,013	70,555	62,174	0,128
16	140	123	– 0,819	69,527	61,146	0,086

Вычислив остатки регрессии e_i (табл. 1.7), по формуле (1.111) находим

$$DW_{\text{набл}} = 0,986,$$

но тогда

$$r \overset{(1.112)}{\approx} 0,507.$$

По статистическим таблицам определяем граничные точки статистики Дарбина–Уотсона

$$d_1 = d_1(0,05;16;1) = 1,1,\ d_2 = d_2(0,05;16;1) = 1,37.$$

Так как $DW_{\text{набл}} \in (0, d_1)$, то гипотеза об отсутствии автокорреляции отвергается.

2) Подействовав на модель (1.109) авторегрессионным преобразованием

$$x_i' = x_i - 0,507\,x_{i-1},\ y_i' = y_i - 0,507\,y_{i-1} \quad (i \in \overline{2,16}),$$

приходим к новому регрессионному уравнению (1.115). Применяя и к нему МНК, находим точечную оценку $\widehat{b}_0' = 6,228$ коэффициента b_0' и ошибки регрессии e_i'.

Для новой модели $DW_{\text{набл}} = 1,71$. Поскольку

$$d_2 = d_2(0,05;15;1) = 1,36,\ DW_{\text{набл}} \in (d_2, 4 - d_2),$$

то гипотеза об отсутствии автокорреляция остатков e_i' принимается. Пересчитаем по формуле (1.114) оценку коэффициента b_0 :

$$\widehat{b}_0 = \frac{6,228}{1 - 0,507} = 12,633.$$

Таким образом, можем записать следующую уточненную оценку исходной модели:

$$\widehat{y} = 12,633 + 0,806\,x.$$

1.5.4. Мультиколлинеарность

Мультиколлинеарность – *взаимозависимость регрессоров* линейной регрессионной множественной модели.

Различают *строгую* и *нестрогую* мультиколлинеарность.

Строгая мультиколлинеарность означает, что $rank\ \mathbf{X} < m + 1$, т.е. не выполнено условие 6^0 Гаусса–Маркова. В этом случае возникают следующие особенности регрессионной модели:

- столбцы матрицы $\mathbf{X}$ линейно зависимы;

• между регрессорами $x_1, \ldots, x_m$ существует линейная функциональная связь;

• матрица $\mathbf{X}^T\mathbf{X}$ вырождена, поэтому $\det\ \mathbf{X}^T\mathbf{X} = 0$.

Поскольку матрица $\mathbf{X}^T\mathbf{X}$ не имеет обратной, основная формула метода наименьших квадратов (1.37) лишена смысла. Поэтому при строгой мультиколлинеарности МНК вообще неприменим.

Однако в практике эконометрического моделирования мультиколлинеарность преимущественно встречается в *нестрогой* форме, когда хотя бы между двумя регрессорами существует сильная корреляционная связь. В этом случае условие 6^0 Гаусса–Маркова формально не нарушается, матрица $\mathbf{X}^T\mathbf{X}$ неособая, но ее определитель крайне мал. Поэтому элементы обратной матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$ становятся очень большими по модулю, но тогда коэффициенты $\widehat{b}_p$ обнаруживают чрезмерно большие стандартные ошибки и зачастую становятся статистически незначимыми.

Из сказанного выше вытекают следующие способы выявления мультиколлинеарности:

1) Анализ корреляционной матрицы между регрессорами (см. формулу (П9))

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{12} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1m} & r_{2m} & \ldots & 1 \end{pmatrix},$$

выявление пар переменных x_i, x_j с большим по модулю коэффициентом корреляции ($|r_{ij}| > 0,8$).

2) Анализ матрицы $\mathbf{X}^T\mathbf{X}$. Признаками мультиколлинеарности являются:

• близость $\det\ \mathbf{X}^T\mathbf{X}$ к нулю;

• плохая обусловленность матрицы $\mathbf{X}^T\mathbf{X}$ (т.е. отношение $\lambda_{\max}/\lambda_{\min}$ велико; здесь $\lambda_{\min}$, $\lambda_{\max}$– наименьшее и наибольшее собственные значения матрицы $\mathbf{X}^T\mathbf{X}$).

Некоторые методы устранения или смягчения мультиколлинеарности

1) Если выделены регрессоры x_i, x_j с большим по модулю коэффициентом корреляции r_{ij}, то один из них может быть удален (обыч-

но исключается тот регрессор, который имеет меньшую экономическую значимость).

2) Применяется *пошаговая процедура отбора наиболее информативных регрессоров.* На первом шаге строятся парные регрессионные модели с каждым регрессором и отбирается тот из них, для которого коэффициент детерминации R^2 будет *наибольшим.* Обозначим выбранный регрессор через x_1. На втором этапе к регрессору x_1 добавляется еще один регрессор x_2, который даст вместе с y, x_1 наибольший коэффициент детерминации, и т.д.

При реализации процедуры отбора на этапах, кроме первого, целесообразно применение *скорректированного* (исправленного) коэффициента детерминации

$$\widehat{R}^2 = 1 - \frac{n-1}{n-m-1} \cdot (1 - R^2). \qquad (1.116)$$

Он в определенной степени исправляет недостаток коэффициента R^2, состоящий в том, что R^2 автоматически увеличивается с ростом числа регрессоров. Отличие $\widehat{R}^2$ от R^2 состоит в добавлении слагаемого $(-m)$ в знаменателе дроби формулы (1.116).

3) Осуществляется переход от регрессоров

$$x_1, \ldots, x_n,$$

связанных сильной корреляционной зависимостью, к новым переменным, которые уже практически не являются коррелированными. Эта, достаточно сложная, процедура проводится в рамках *метода главных компонент,* применяемого в многомерном статистическом анализе.

Контрольные вопросы и задания

1. Запишите общий вид уравнения регрессии, укажите смысл слагаемых его правой части.

2. Запишите общий вид регрессионной модели в наблюдениях.

3. В чем состоит отличие моделей парной регрессии и множественной регрессии?

4. В каких случаях целесообразно примененение парной регрессионной модели?

5. Постройте на плоскости корреляционное поле, которое, на ваш взгляд, допускает применение парной линейной регрессионной модели. Постройте другое регрессионное поле, когда применение такой модели, по вашему мнению, невозможно.

6. В чем отличие ошибок регрессии ε_i и остатков регрессии e_i?

7. Укажите геометрический смысл метода наименьших квадратов для оценки коэффициентов парной линейной регрессии.

8. Запишите матричную форму модели линейной регрессии, укажите вид входящих в нее матриц.

9. Объясните, почему первый столбец регрессионной матрицы $\mathbf{X}$ состоит из единиц.

10. Запишите основное в методе наименьших квадратов матричное уравнение, определяющее вектор-столбец $\widehat{\mathbf{b}}$ точечных оценок коэффициентов линейной регрессии.

11. Запишите матричное уравнение, определяющее вектор-столбец $\widehat{\mathbf{b}}_r$ точечных оценок коэффициентов при регрессорах, охарактеризуйте входящие в него матрицы.

12. Докажите вторую и третью из формул (1.55).

13. Исходя из 3-го свойства МНК-оценок, докажите, что среднее значение остатков регрессии

$$\overline{e} = \frac{\sum\limits_{i=1}^{n} e_i}{n}$$

в модели линейной регрессии (1.30) равно нулю.

14. Проверьте, что остатки e_i линейных регрессий, построенных в примерах 1 и 2, удовлетворяют формуле (1.57).

15. Сформулируйте условия Гаусса–Маркова классической линейной регрессионной модели и теорему Гаусса–Маркова.

16. Кратко охарактеризуйте основные этапы проверки статистического качества линейной регрессионной модели.

2. Нелинейные регрессии

Рассмотренная в первой главе теория линейных регрессионных моделей составляет ядро эконометрики. Напомним некоторые особенности линейных моделей:

• система нормальных уравнений, определяющая МНК-оценки коэффициентов регрессии, является линейной, что позволяет для ее решения использовать эффективный аппарат линейной алгебры;

• для классической линейной модели разработаны удобные алгоритмы исследования ее статистических свойств;

• коэффициенты регрессии допускают экономическую интерпретацию.

Однако ограничиться лишь линейными моделями нельзя, поскольку экономические процессы зачастую описываются нелинейными соотношениями (например, производственными функциями Кобба–Дугласа, см. п. 2.2). Поэтому возникает необходимость изучать регрессионные модели (1.2) с *нелинейными* функциями регрессии (1.3). Мы убедимся, что во многих важных случаях такие нелинейные модели удается так или иначе *линеаризовать* — свести к линейным регрессионным моделям.

В основном будем рассматривать *парную нелинейную регрессию*

$$y = f(x) + \varepsilon. \tag{2.1}$$

Функция регрессии $f(x)$ определяет структуру модели; она содержит параметры, которые необходимо оценить по выборке (1.5).

Парная нелинейная регрессионная модель (2.1) полностью определяется функцией регрессии $f(x)$. Поэтому классификация таких моделей сводится к классификации их функций регрессии.

Различают два класса нелинейных регрессий:

1) *квазилинейные* регрессии, линейные по оцениваемым параметрам и нелинейные по аргументу x :

$$f(x) = b_0 + b_1 a_1(x) + \ldots + b_m a_m(x) \quad (m \in \mathbb{N}), \tag{2.2}$$

здесь

$$a_1(x), \ldots, a_m(x)$$

суть заданные функции (вообще говоря, нелинейные);

2) регрессии, нелинейные и по параметрам, и по аргументу.

2.1. Квазилинейные регрессии

Назовем *базовые,* часто применяемые квазилинейные регрессии:

- полиномиальная регрессия

$$f(x) = b_0 + b_1 x + \ldots + b_m x^m; \tag{2.3}$$

- логарифмическая регрессия

$$f(x) = b_0 + b_1 \ln x; \tag{2.4}$$

- гиперболическая регрессия

$$f(x) = b_0 + \frac{b_1}{x}\,. \tag{2.5}$$

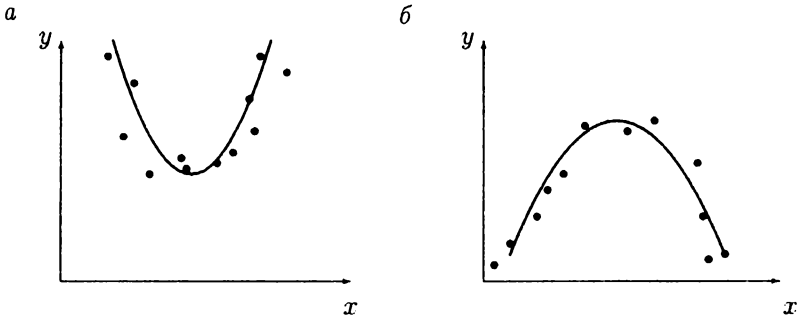


Рис. 2.1. Квадратичная регрессия:
а – случай $b_2 > 0$; *б* – случай $b_2 < 0$

Качество квазилинейной модели во многом определяется удачным выбором функции $f(x)$, осуществляемым на этапе *спецификации* модели.

Дадим некоторые рекомендации по применению базовых регрессий.

Из полиномиальных нелинейных регрессий (2.3) ($m > 1$) чаще всего применяется *квадратичная*, или *параболическая*, *регрессия* ($m = 2$). Ее целесообразно использовать при *изменении* характера связи между переменными x, y : возрастания на убывание или наоборот (рис. 2.1).

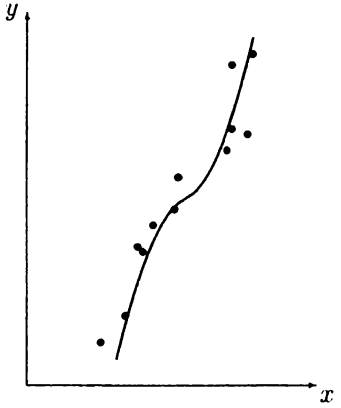


Рис. 2.2. Кубическая регрессия

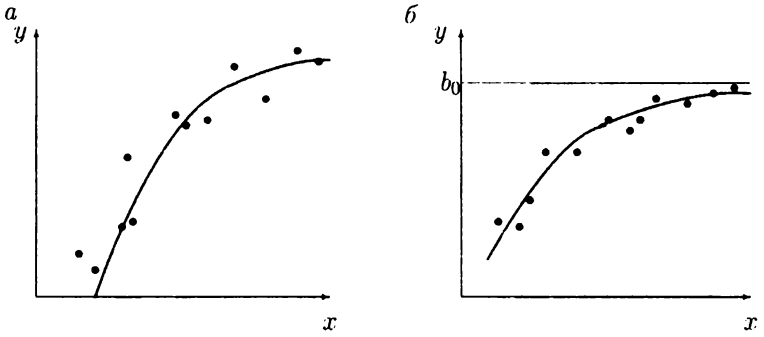


Рис. 2.3. Базовые регрессии:
а – логарифмическая; *б* – гиперболическая

Кубическую регрессию ($m = 3$) можно применять при монотонности нелинейного изменения y относительно x (рис. 2.2). Предполагается, что кубическая функция регрессии монотонна.

Если же с ростом x переменная y неограниченно растет, но медленнее, чем в случае полиномиальной регрессии, то возможно применение логарифмической регрессии (2.4).

Гиперболическую регрессию (2.5) используют, когда с ростом x значения переменной y в среднем стабилизируются относительно некоторой величины b_0. Кривая регрессии $y = b_0 + b_1/x$ имеет горизонтальную асимптоту $y = b_0$, поскольку $\lim\limits_{x\to\infty}(b_0 + b_1/x) = b_0$ (рис. 2.3).

Квазилинейная регрессионная модель (2.1), (2.2) *введением новых переменных линеаризуется*, что и оправдывает ее название. Действительно, полагая

$$x_1 = a_1(x), \ldots, x_m = a_m(x), \tag{2.6}$$

преобразуем квазилинейную модель к виду

$$y = b_0 + b_1 x_1 + \ldots + b_m x_m + \varepsilon.$$

Этот подход при $m = 1$ приводит к парной линейной регрессии. Таким способом, в частности, изучают логарифмическую и гиперболическую регрессии.

В случае $m > 1$ приходим к множественной линейной регрессии, причем удобно *явно не вводить* новые переменные $x_1, \ldots, x_m$. Запишем модель в наблюдениях, отвечающую исходной квазилинейной модели (2.1), (2.2):

$$y_i = b_0 + b_1 a_1(x_i) + \ldots + b_m a_m(x_i) + \varepsilon_i \quad (i \in \overline{1,n}). \tag{2.7}$$

Представим модель (2.7) в матричной форме:

$$\mathbf{y} = \mathbf{A}\mathbf{b} + \boldsymbol{\varepsilon}. \tag{2.8}$$

Здесь используются те же обозначения, что и в п. 1.4, только вместо регрессионной матрицы $\mathbf{X}$ появляется матрица

$$\mathbf{A} = \begin{pmatrix} 1 & a_1(x_1) & a_2(x_1) & \ldots & a_m(x_1) \\ 1 & a_1(x_2) & a_2(x_2) & \ldots & a_m(x_2) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & a_1(x_n) & a_2(x_n) & \ldots & a_m(x_n) \end{pmatrix}. \tag{2.9}$$

Рассматриваем *классическую квазилинейную модель*, т.е. предполагаем, что для модели (2.8) выполнены условия Гаусса–Маркова, приведенные в подп. 1.3.1 (с заменой $\mathbf{X}$ на $\mathbf{A}$).

Исследование модели проводится по общей схеме, изложенной в подп. 1.3.2-1.3.6 (с учетом обозначений (2.6)). Однако двумерность исходной квазилинейной модели позволяет дать наглядную интерпретацию постановки задачи и результатов моделирования.

Методом наименьших квадратов находится вектор $\widehat{\mathbf{b}}$ точечных оценок коэффициентов регрессии

$$\widehat{\mathbf{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}. \tag{2.10}$$

Он определяет на плоскости (x, y) линию регрессии

$$y = \widehat{f}(x) = \widehat{b}_0 + \widehat{b}_1 a_1(x) + \ldots + \widehat{b}_m a_m(x). \tag{2.11}$$

Выявление соответствия модели квазилинейной регрессии наблюдениям в целом проводится в рамках проверки гипотезы

$$H_0 : \; b_1 = \ldots = b_m = 0.$$

Для этого используются коэффициент детерминации R^2 и F- статистика.

Вводятся стандартная ошибка регрессии

$$s = \frac{\sum\limits_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n-m-1} = \sqrt{\frac{nk_{yy}(1-R^2)}{n-m-1}}$$

и стандартные ошибки коэффициентов регрессии $\widehat{b}_l$ $(l \in \overline{0.m})$

$$s_l = s\sqrt{((\mathbf{A}^T\mathbf{A})^{-1})_{ll}}.$$

Они позволяют определить доверительные интервалы истинных коэффициентов регрессии b_l

$$I_\gamma(b_l) = (\widehat{b}_l - \delta_l, \widehat{b}_l + \delta_l), \; \delta_l = s_l \cdot t_{1-\alpha/2}(n-m-1),$$

отвечающие надежности $\gamma = 1 - \alpha$.

Укажем расчетные формулы, позволяющие исследовать модель *квадратичной* регрессии, т.е. модель (2.1) с функцией регрессии

$$f(x) = b_0 + b_1 x + b_2 x^2.$$

Теперь $a_1(x) = x$, $a_2(x) = x^2$, поэтому эти формулы получатся из формул (1.49), (1.54), (1.79), (1.95) заменами: x_1 (или 1 – в обозначении ковариаций) на x, x_2 (или 2) на x^2.

Находятся ковариации между величинами x, x^2, y :

$$k_{xx} = \overline{x^2} - (\overline{x})^2,\ k_{xx^2} = \overline{x^3} - \overline{x} \cdot \overline{x^2},\ k_{x^2x^2} = \overline{x^4} - (\overline{x^2})^2,$$

$$k_{xy} = \overline{xy} - \overline{x} \cdot \overline{y},\ k_{x^2y} = \overline{x^2y} - \overline{x^2} \cdot \overline{y},\ k_{yy} = \overline{y^2} - (\overline{y})^2.$$

Точечные оценки параметров регрессии определяются по формулам

$$\widehat{b}_1 = \frac{k_{x^2x^2} \cdot k_{xy} - k_{xx^2} \cdot k_{x^2y}}{\Delta},\ \widehat{b}_2 = \frac{k_{xx} \cdot k_{x^2y} - k_{xx^2} \cdot k_{xy}}{\Delta},\ \widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x} - \widehat{b}_2\overline{x^2}.$$

Здесь

$$\Delta = k_{xx} \cdot k_{x^2x^2} - (k_{xx^2})^2.$$

Коэффициент детерминации находится следующим образом:

$$R^2 = \frac{\widehat{b}_1 k_{xy} + \widehat{b}_2 k_{x^2y}}{k_{yy}}.$$

Наконец, стандартные ошибки коэффициентов регрессии определяются так:

$$s_0 = s \cdot \sqrt{\frac{\overline{x^2} \cdot \overline{x^4} - (\overline{x^3})^2}{n\Delta}},\ s_1 = s \cdot \sqrt{\frac{k_{x^2x^2}}{n\Delta}},\ s_2 = s \cdot \sqrt{\frac{k_{xx}}{n\Delta}}.$$

Для построения кривой квадратичной регрессии – параболы

$$y = \widehat{f}(x) = \widehat{b}_0 + \widehat{b}_1 x + \widehat{b}_2 x^2$$

целесообразно найти ее вершину

$$(x_0, \widehat{f}(x_0)).$$

Здесь $x_0 = -\widehat{b}_1/(2\widehat{b}_2)$ – критическая точка функции $\widehat{f}(x)$, т.е. такая, что $\widehat{f}'(x_0) = 0$. Следует учесть, что парабола симметрична относительно прямой $x = x_0$:

$$\widehat{f}(x_0 - c) = \widehat{f}(x_0 + c).$$

Пример. Анализируется прибыль предприятия y (млн у.е.) в зависимости от его расходов на рекламу x (млн у.е.). Известны следующие статистические данные (табл. 2.1)[1].

[1]См. работу С.А. Бородича [3, с. 226].

Таблица 2.1

x_i	0,8	1	1,8	2,5	4	5,7	7,5	8,3	8,8
y_i	5	7	13	15	20	25	22	20	17

Требуется:

1) По характеру корреляционного поля на плоскости (x, y) убедиться, что целесообразно применение квадратичной регрессии.

2) Найти точечные оценки коэффициентов регрессии и построить на плоскости линию регрессии.

3) Применяя коэффициент детерминации R^2, на уровне значимости $\alpha = 0,05$ проверить гипотезу о соответствии модели наблюдениям.

4) Если модель соответствует наблюдениям, то с надежностью $\gamma = 0,95$ найти доверительные интервалы коэффициентов регрессии.

Решение:

1) Корреляционное поле построено на рис. 2.4. Рисунок показывает, что с ростом значений x_i величины y_i вначале растут, а затем уменьшаются. Возможно применение квадратичной регрессии.

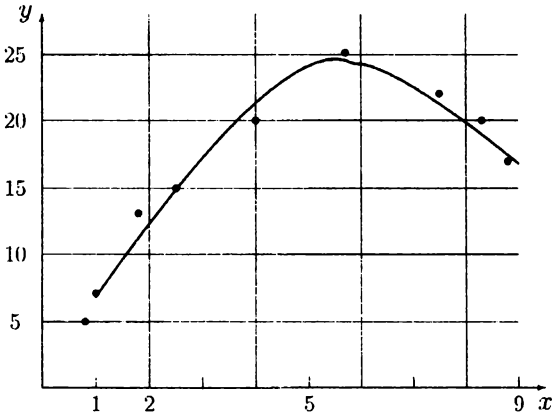


Рис. 2.4. Квадратичная регрессия. Пример п. 2.1

2) Для нахождения средних величин заполним табл. 2.2.

После этого вычислим выборочные ковариации:

$$k_{xx} = \overline{x^2} - (\overline{x})^2 = 29,133 - 4,489^2 = 8,982;$$

$$k_{x^2x^2} = \overline{x^4} - (\overline{x^2})^2 = 1696,602 - 29,133^2 = 847,87;$$

$$k_{xx^2} = \overline{x^3} - \overline{x} \cdot \overline{x^2} = 216,366 - 4,489 \cdot 29,133 = 85,588;$$

$$k_{xy} = \overline{xy} - \overline{x} \cdot \overline{y} = 86,111 - 4,489 \cdot 16 = 14,287;$$

$$k_{x^2y} = \overline{x^2y} - \overline{x^2} \cdot \overline{y} = 578,9 - 29,133 \cdot 16 = 112,772;$$

$$k_{yy} = \overline{y^2} - (\overline{y})^2 = 296,222 - 16^2 = 40,222.$$

Найдем

$$\Delta = k_{xx} \cdot k_{x^2x^2} - k_{xx^2}^2 = 8,982 \cdot 847,87 - 85,588^2 = 290,262.$$

Таблица 2.2

i	x_i	x_i^2	x_i^3	x_i^4	y_i	y_i^2	x_iy_i	$x_i^2y_i$
1	0,8	0,64	0,512	0,41	5	25	4	3,2
2	1	1	1	1	7	49	7	7
3	1,8	3,24	5,83	10,5	13	169	23,4	42,12
4	2,5	6,25	15,62	39,06	15	225	37,5	93,75
5	4	16	64	256	20	400	80	320
6	5,7	32,49	185,19	1055,6	25	625	142,5	812,25
7	7,5	56,25	421,88	3164,06	22	484	165	1237,5
8	8,3	68,89	571,79	4745,83	20	400	166	1377,8
9	8,8	77,44	681,47	5996,95	17	289	149,6	1316,5
$\sum$	40,4	262,2	1947,3	15269,42	144	2666	775	5210,1
$\sum/n$	4,489	29,133	216,366	1696,602	16	296,222	86,111	578,9
Среднее	$\overline{x}$	$\overline{x^2}$	$\overline{x^3}$	$\overline{x^4}$	$\overline{y}$	$\overline{y^2}$	$\overline{xy}$	$\overline{x^2y}$

Вычисляем точечные оценки коэффициентов регрессии:

$$\widehat{b}_1 = \frac{k_{x^2x^2} \cdot k_{xy} - k_{xx^2} \cdot k_{x^2y}}{\Delta} = \frac{847,87 \cdot 14,287 - 85,588 \cdot 112,772}{290,262} =$$

$$= \frac{12113,518 - 9651,93}{290,262} = \frac{2461,588}{290,262} = 8,48;$$

$$\widehat{b}_2 = \frac{k_{xx} \cdot k_{x^2y} - k_{xx^2} \cdot k_{xy}}{\Delta} = \frac{8,982 \cdot 112,772 - 85,588 \cdot 14,287}{290,262} =$$

$$= \frac{1012,918 - 1222,796}{290,262} = \frac{-209,878}{290,262} = -0,723;$$

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x} - \widehat{b}_2\overline{x^2} = 16 - 8,479 \cdot 4,489 + 0,723 \cdot 29,133 =$$
$$= 16 - 38,062 + 21,063 = -0,999.$$

Тем самым определено уравнение кривой регрессии параболы

$$y = -0,999 + 8,48\,x - 0,723\,x^2.$$

Найдем точку x_0 – абсциссу вершины параболы:

$$x_0 = -\frac{\widehat{b}_1}{2\widehat{b}_2} = \frac{8,479}{2 \cdot 0,723} = 5,86.$$

Для построения параболы вычислим некоторые ее точки, включая вершину (табл. 2.3)[2].

Таблица 2.3

x	1	2	3	4	5	5,87	6	7	8	9
$\widehat{f}(x)$	6,75	13,06	17,92	21,33	23,3	23,84	23,83	22,91	20,55	16,75

3) Для проверки соответствия модели результатам наблюдений найдем коэффициент детерминации

$$R^2 = \frac{\widehat{b}_1 k_{xy} + \widehat{b}_2 k_{x^2y}}{k_{yy}} = \frac{8,48 \cdot 14,287 - 0,723 \cdot 112,772}{40,222} =$$
$$= \frac{121,154 - 81,534}{40,222} = \frac{39,62}{40,222} = 0,985.$$

Вычислим наблюдаемое значение статистики Фишера

$$F_{\text{набл}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = [n = 9, m = 2] = \frac{3 \cdot 0,985}{0,015} = 197.$$

Выбрав уровень значимости $\alpha = 0,05$, по таблицам распределения Фишера находим

$$F_{\text{крит}} = F_{1-\alpha}(m, n - m - 1) = F_{0,95}(2, 6) = 5,14 < F_{\text{набл}}.$$

[2]Мы ограничиваемся в таблице двумя знаками после запятой, поскольку для построения параболы эта точность достаточна.

Следовательно, построенная квадратичная регрессионная модель на уровне значимости $\alpha = 0,05$ соответствует наблюдениям.

4) Для нахождения доверительных интервалов параметров регрессии b_0, b_1, b_2 предварительно вычислим стандартные ошибки их точечных оценок:

$$s_0 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot (\overline{x^2} \cdot \overline{x^4} - (\overline{x^3})^2)} =$$

$$= \sqrt{\frac{40,222 \cdot 0,015}{6 \cdot 290,262} \cdot (29,133 \cdot 1696,602 - 216,366^2)} =$$

$$= \sqrt{0,00035 \cdot 2612,861} = \sqrt{0,914} = 0,956;$$

$$s_1 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot k_{x^2x^2}} = \sqrt{0,00035 \cdot 847,87} = \sqrt{0,297} = 0,545;$$

$$s_2 = \sqrt{\frac{k_{yy}(1-R^2)}{(n-3)\Delta} \cdot k_{xx}} = \sqrt{0,00035 \cdot 8,982} = \sqrt{0,00314} = 0,056.$$

По таблицам распределения Стьюдента

$$t_{1-\alpha/2}(n-3) = t_{0,975}(6) = 2,45.$$

Находим

$$\delta_0 = s_0 \cdot t_{1-\alpha/2}(n-3) = 0,956 \cdot 2,45 = 2,342;$$

$$\delta_1 = s_1 \cdot t_{1-\alpha/2}(n-3) = 0,545 \cdot 2,45 = 1,335;$$

$$\delta_2 = s_2 \cdot t_{1-\alpha/2}(n-3) = 0,056 \cdot 2,45 = 0,137.$$

Поэтому

$$I_{0,95}(b_0) = (\widehat{b}_0 - \delta_0,\ \widehat{b}_0 + \delta_0) = (-0,999 - 2,342;\ -0,999 + 2,342) =$$

$$= (-3,341;\ 1,343),$$

$$I_{0,95}(b_1) = (\widehat{b}_1 - \delta_1,\ \widehat{b}_1 + \delta_1) = (8,48 - 1,335;\ 8,48 + 1,335) =$$

$$= (7,145;\ 9,815),$$

$$I_{0,95}(b_2) = (\widehat{b}_2 - \delta_2,\ \widehat{b}_2 + \delta_2) = (-0,723 - 0,137;\ -0,723 + 0,137) =$$

$$= (-0,86;\ -0,586).$$

2.2. Нелинейные по параметрам регрессии и их линеаризации

Нелинейные по параметрам регрессии разбиваются на две группы:

1) *внутренне линейные регрессии*, допускающие линеаризацию (сведение к линейной регрессии) с помощью подходящих подстановок переменных;

2) *внутренне нелинейные регрессии*, для которых нет удобных способов линеаризации (например, $y = b_0 + b_1 e^{b_2 x} + \varepsilon$).

1. Укажем некоторые базовые внутренне линейные парные регрессии:

- степенная регрессия

$$y = b_0 x^{b_1} \varepsilon; \tag{2.12}$$

- показательная (экспоненциальная) регрессия

$$y = b_0 e^{b_1 x + \varepsilon}; \tag{2.13}$$

- логистическая регрессия

$$y = \frac{1}{b_0 + b_1 e^{-x} + \varepsilon} \quad (b_1 > 0). \tag{2.14}$$

Обращаем внимание на различные способы введения в уравнения (2.12) (2.14) случайной ошибки ε. В формулу (2.12) она входит множителем (мультипликативно), в формулу (2.13) – слагаемым (аддитивно) в показателе степени, в формулу (2.14) – слагаемым в знаменателе. Такие способы позволяют линеаризовать регрессии (2.12) (2.14).

Степенная регрессия применяется при анализе зависимости объема y продукции от объема x затрат основного ресурса.

Показательная регрессия (2.13) позволяет моделировать быстро растущие экономические объекты с постоянным темпом относительного роста. Действительно, дифференцируя равенство (2.13) по x, находим

$$y'_x = b_0 b_1 e^{b_1 x + \varepsilon} = b_1 y,$$

так что темп относительного роста

$$\frac{y'_x}{y} = b_1 = const.$$

Логистическая регрессия используется в случаях зависимости y от x вида

$$y = \frac{1}{b_0 + b_1 e^{-x}} \tag{2.15}$$

(формула (2.15) получена из (2.14) при $\varepsilon = 0$). Функция (2.15) возрастает, причем

$$0 < y < \frac{1}{b_0}, \quad \lim_{x \to -\infty} y = 0, \quad \lim_{x \to +\infty} y = \frac{1}{b_0}$$

(рис. 2.5).

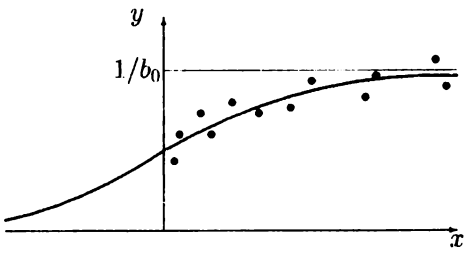


Рис. 2.5. Логистическая регрессия

Степенная и показательная регрессии линеаризуются с помощью логарифмирования и подстановок. Подробнее рассмотрим степенную регрессию. Логарифмируя равенство (2.12) по основанию e, получаем соотношение

$$\ln y = \ln b_0 + b_1 \ln x + \ln \varepsilon.$$

Замены

$$y' = \ln y, \; x' = \ln x$$

приводят к линейной регрессии

$$y' = \ln b_0 + b_1 x' + \ln \varepsilon.$$

Исследуя ее, находят методом наименьших квадратов точечные оценки параметров $\ln b_0$, b_1

$$\widehat{\ln b_0}, \; \widehat{b_1}$$

и изучают их статистические свойства. Точечная оценка исходного параметра b_0 определяется так:

$$\widehat{b_0} = e^{\widehat{\ln b_0}}.$$

Аналогично исследуется множественная степенная регрессия

$$y = b_0 \cdot x_1^{b_1} \cdot \ldots \cdot x_m^{b_m} \cdot \varepsilon \quad (b_i > 0). \tag{2.16}$$

Она определяется функцией

$$y = b_0 x_1^{b_1} x_2^{b_2} \cdot \ldots \cdot x_m^{b_m}, \tag{2.17}$$

которую называют *производственной функцией Кобба-Дугласа*. Такие функции выражают зависимость объема y выпуска некоторой продукции от объемов $x_1, \ldots, x_m$ используемых при производстве этой продукции ресурсов. Поэтому множественная степенная регрессия (2.16) применяется при анализе зависимостей, определяемых производственными функциями нескольких переменных (2.17). Логарифмированием она приводится к линейной модели

$$\ln y = \ln b_0 + b_1 \ln x_1 + \ldots + b_m \ln x_m + \ln \varepsilon.$$

Уравнение логистической регрессии (2.14) подстановками

$$y' = \frac{1}{y}\,, \; x' = e^{-x}$$

приводится к линейной регрессии

$$y' = b_0 + b_1 x' + \varepsilon.$$

При исследовании внутренне линейных моделей обычно предполагается, что их линеаризации являются классическими линейными регрессионными моделями.

2. Так как аналитическое исследование *внутренне нелинейных регрессионных моделей* затруднено, для этих целей обычно используются компьютерные пакеты (Excel, Mathcad, Statistica, Statgraphics, Econometric Views и др.)[3]. Нормальные уравнения системы, определяющей точечные оценки коэффициентов регрессии, становятся *нелинейными*, что усложняет анализ внутренне нелинейных моделей. Для решения системы нормальных уравнений применяются различные численные методы локальной минимизации, требующие задания начальных приближений коэффициентов. Статистический анализ моделей проводится с

[3] Excel, Mathcad – универсальные математические пакеты, Statistica, Statgraphics – статистические пакеты, Econometric Views – специализированный эконометрический пакет.

помощью приближенных схем проверки гипотез и нахождения доверительных интервалов.

Компьютерные пакеты широко применяются в эконометрической практике и при исследовании линейных и линеаризуемых регрессионных моделей, временных рядов и систем одновременных уравнений. Особенности работы с этими пакетами описаны в учебной литературе [2, с. 46–82; 12, с. 279–287; 14, с. 542–546].

Контрольные вопросы и задания

1. Какие из приведенных ниже нелинейных функций регрессии линейны по параметрам? Какие функции нелинейны по параметрам?

$$f(x) = b_0 + b_1 x + b_2 x^2;$$

$$f(x) = e^{b_0 x}(b_1 x + b_2 x^2);$$

$$f(x) = b_0 + \frac{b_1}{x} + b_2 \ln x;$$

$$f(x) = \frac{b_0 + b_1 x}{b_2 + b_3 x}.$$

2. Запишите формулы нахождения МНК-оценок параметров нелинейной регрессии (2.1) в случаях логарифмической функции регрессии (2.4) и гиперболической функции регрессии (2.5).

3. Укажите способ линеаризации показательной регрессии (2.13).

3. Временные ряды

Применим развитый в предыдущих главах аппарат регрессионного анализа к *временны́м рядам*, связанным с *временны́ми статистическими данными*. Временные ряды моделируют *экономические процессы*, описывая их изменение во времени.

Вначале будем рассматривать *одномерные временные ряды*. Так называют конечную последовательность, состоящую из n наблюдений над признаком (случайной величиной) y в последовательные равноотстоящие моменты времени с шагом T. Изменяя начало переменной «время» и масштабируя ее, можем добиться, чтобы наблюдаемые моменты времени определялись натуральными числами

$$1, 2, \ldots, n. \tag{3.1}$$

Тогда временной ряд запишется следующим образом:

$$\{y_t\} \quad (t \in \overline{1,n}). \tag{3.2}$$

Значения (уровни) y_t временного ряда (3.2) формируются под влиянием нескольких компонент. Будем рассматривать *аддитивную модель* временного ряда, где эти компоненты *суммируются*:

$$y_t = f_t + \alpha_t + \beta_t + \varepsilon_t \quad (t \in \overline{1,n}). \tag{3.3}$$

Здесь

- f_t — *тренд*, т.е. основная тенденция развития временного ряда, описывающая устойчивую динамику экономического процесса в течение длительного времени;
- α_t - *сезонная компонента*, отражающая повторяемость экономического процесса с относительно небольшим периодом (неделя, месяц, год);
- β_t — *циклическая компонента*, отражающая повторяемость с длительными периодами (например, волны экономической активности Кондратьева, циклы солнечной активности);

• ε_t – *случайная компонента.*

Величины f_t, α_t, β_t являются детерминированными, неслучайными. Они не обязательно присутствуют в модели (3.3) временного ряда. Например, если шаг по времени T равен одному году, то модель не улавливает сезонных изменений и $\alpha_t = 0$. Временные ряды, описывающие динамику постсоветской России, могут не содержать циклической компоненты в силу небольшого времени наблюдений, и тогда $\beta_t = 0$.

По структуре модель (3.3) является моделью парной регрессии, где *роль регрессора играет время* t. При этом исходный ряд (3.2), описывающий наблюдения над изучаемым признаком y, задает временную выборку

$$(t, y_t) \quad (t \in \overline{1,n}). \tag{3.4}$$

Ее можно трактовать геометрически как корреляционное поле на плоскости (t, y). Это поле имеет следующую особенность: абсциссы его точек пробегают множество (3.1). Для наглядности соседние точки поля можно соединить отрезками прямых. Полученную ломаную называют графиком (исходного, несглаженного) временного ряда (рис. 3.1).

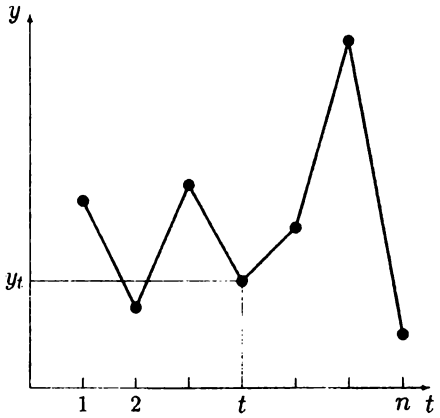


Рис. 3.1. График временного ряда

Принципиальное отличие временного ряда от рассматриваемых ранее классических регрессионных моделей состоит в том, что в большинстве случаев различные наблюдения y_t *не являются независимыми и одинаково распределенными* случайными величинами. При исследова-

нии временного ряда *изучается характер этой зависимости*, что позволяет решать задачи анализа и прогнозирования экономических процессов.

Отметим *основные этапы исследования временных рядов*:

1) выделение неслучайных компонент временного ряда;

2) исследование случайной составляющей временного ряда;

3) прогнозирование развития изучаемого процесса на будущее время;

4) исследование взаимосвязей между различными одномерными временными рядами.

3.1. Методы выделения неслучайной составляющей временного ряда

Выделение неслучайной составляющей

$$f_t + \alpha_t + \beta_t$$

является сглаживанием (выравниванием) временного ряда (3.3).

Остановимся на методах выделения тренда — функции f_t. Они делятся на *аналитические (регрессионные)* и *алгоритмические (методы скользящего среднего)*.

1. В рамках *аналитических методов*, как и при регрессионном анализе, 1) вначале выбирают *структуру* модели, т.е. общий вид функции тренда f_t, а затем 2) по выборке (3.4) оценивают параметры этой функции.

1) В качестве функций структуры f_t обычно выбирают те же функции, что и в случае парной (вообще говоря, нелинейной) регрессии (см. гл. 2). Наиболее часто используются следующие функции:

- полиномиальная

$$f_t = b_0 + b_1 t + \ldots + b_m t^m \quad (m \in \mathbb{N}) \tag{3.5}$$

(при $m = 1$ — линейная);

- экспоненциальная

$$f_t = b_0 e^{b_1 t};$$

- логистическая

$$f_t = \frac{1}{b_0 + b_1 e^{-t}}.$$

Если выбрана полиномиальная функция тренда – многочлен (3.5), то для нахождения степени m этого многочлена либо а) учитывают характер корреляционного поля, либо б) вычисляют *разности* уровней ряда y_t *возрастающих порядков*: разности первого порядка $\Delta y_t = y_t - y_{t-1}$, разности второго порядка $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$ и т.д. Наименьший порядок m, при котором разности равны или близки, и принимают за степень многочлена.

2) Для оценивания параметров функции f_t выбранной структуры применяют метод наименьших квадратов. Пусть рассматривается многочлен первой степени – линейная функция

$$f_t = b_0 + b_1 t.$$

Тогда модель (3.3) предстает в виде парной линейной регрессионной модели

$$y_t = b_0 + b_1 t + \varepsilon_t. \tag{3.6}$$

Если модель (3.6) удовлетворяет условиям Гаусса–Маркова, то для оценки ее коэффициентов применяется обыкновенный МНК. Если остатки ε_t взаимно коррелированы, то – ОМНК.

В случае *нелинейной* функции f_t оценивание ее коэффициентов проводится в рамках нелинейного парного регрессионного моделирования.

Пример 3.1. Заданы объемы перевозок (в тыс. т) по автотранспортному предприятию в 1993–2000 годах (табл. 3.1)[1].

Таблица 3.1

Год	1993	1994	1995	1996	1997	1998	1999	2000
y	360	400	401	422	430	463	485	505

Требуется:

1) Построить на плоскости (t, y) график исходного временного ряда. Для применения полиномиальной регрессии определить степень многочлена (3.5).

2) С помощью МНК найти точечные оценки коэффициентов регрессии. Построить на плоскости (t, y) линию регрессии.

3) На уровне значимости $\alpha = 0,05$ проверить соответствие модели наблюдениям.

[1]См. работу М.Р. Ефимовой и др. [8, с. 241].

4) В случае соответствия вычислить точечный прогноз объема перевозок на 2002 год.

Решение:

1) Произведем сдвиг по времени. Году $1992 + t$ сопоставим натуральное число t, при этом $1 \leqslant t \leqslant 8$. Построим на плоскости (t, y) корреляционное поле. Соединив его соседние точки отрезками прямых, получим график исходного временного ряда (рис. 3.2). Поскольку он близок к прямой, выберем *линейный* тренд $f_t = b_0 + b_1 t$.

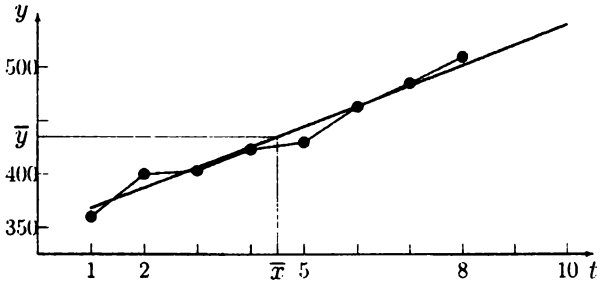


Рис. 3.2. Линейный тренд. Пример 3.1

2) Найдем точечные оценки коэффициентов модели (3.6). Для отыскания средних величин

$$\overline{t},\ \overline{t^2},\ \overline{y},\ \overline{y^2},\ \overline{ty}$$

(они вычисляются по формулам (1.19), но с заменой x, x_i, i на t) заполним расчетную табл. 3.2. После этого найдем ковариации

$$k_{tt} = \overline{t^2} - (\overline{t})^2 = 25,5 - (4,5)^2 = 5,25;$$

$$k_{ty} = \overline{ty} - \overline{t} \cdot \overline{y} = 2051,75 - 4,5 \cdot 433,25 = 102,125;$$

$$k_{yy} = \overline{y^2} - (\overline{y})^2 = 189750,5 - (433,25)^2 = 2044,94.$$

Теперь вычислим

$$\widehat{b}_1 = \frac{k_{ty}}{k_{tt}} = \frac{102,125}{5,25} = 19,452, \quad \widehat{b}_0 = \overline{y} - \overline{t} \cdot \widehat{b}_1 = 433,25 - 4,5 \cdot 19,452 = 345,716.$$

Запишем оцененную функцию линейного тренда

$$\widehat{y}_t = 345,716 + 19,452 \cdot t.$$

Таблица 3.2

–	t	t^2	y_t	y_t^2	ty_t
–	1	1	360	129600	360
–	2	4	400	160000	800
–	3	9	401	160801	1203
–	4	16	422	178084	1688
–	5	25	430	184900	2150
–	6	36	463	214369	2778
–	7	49	485	235225	3395
–	8	64	505	255025	4040
$\sum$	36	204	3466	1518004	16414
$\sum/n$	4,5	25,5	433,25	189750,5	2051,75
Средние	$\overline{t}$	$\overline{t^2}$	$\overline{y}$	$\overline{y^2}$	$\overline{ty}$

Построим на том же рисунке график этой функции. Это и будет прямая регрессии. Она выравнивает график исходного временного ряда.

3) Вычислим коэффициент детерминации

$$R^2 \overset{(1.81)}{=} \frac{k_{ty}^2}{k_{tt} \cdot k_{yy}} = \frac{(102,125)^2}{5,25 \cdot 2044,94} = \frac{10429,515}{10735,935} = 0,971.$$

Применив F-статистику (1.82) при $m = 1$, вычислим

$$F_{\text{набл}} = \frac{0,971 \cdot 6}{0,029} = \frac{5,826}{0,029} = 200,9.$$

Найдем по таблицам квантилей распределения Фишера

$$F_{\text{крит}} = F_{1-\alpha}(1, n-2) = F_{0,95}(1,6) = 5,99 < F_{\text{набл}}.$$

Поскольку $F_{\text{крит}} < F_{\text{набл}}$, построенная модель соответствует наблюдениям.

4) Для нахождения точечного прогноза на 2002 год вычисляем

$$\widehat{y}_{10} = 345,716 + 19,452 \cdot 10 = 540,236 \text{ (тыс. т)}.$$

2. Кратко остановимся на *алгоритмических методах* выравнивания временных рядов (методах скользящего среднего) [12, с. 143–144].

Они основаны на переходе от исходных значений y_t временного ряда к их средним значениям $\widetilde{y}_t$ на всех отрезках времени фиксированной длины m. При этом отрезок длиной m скользит по оси времени. Полученный ряд скользящих средних $\{\widetilde{y}_t\}$ ведет себя более гладко, чем исходный, поскольку дисперсия значений нового ряда $D(\widetilde{y}_t)$ меньше, чем дисперсия значений исходного временного ряда $D(y_t)$.

В качестве средних значений часто выбирают средние арифметические, возможно с весами.

Напомним определение взвешенного среднего арифметического. Рассмотрим группу чисел

$$y_1,\ y_2,\ \ldots,\ y_m. \tag{3.7}$$

Введем их веса $w_1,\ w_2,\ \ldots,\ w_m$ такие, что $w_i > 0,\ \sum\limits_{i=1}^{m} w_i = 1$. Взвешенное среднее арифметическое чисел (3.7) определяется формулой

$$\widetilde{y}_{взв} = \frac{\sum\limits_{i=1}^{m} w_i\, y_i}{m}. \tag{3.8}$$

Разные числа (3.7) вносят неодинаковый вклад в величину (3.8). Иногда требуется, чтобы основную роль играли *последние* числа, тогда их веса выбирают бо́льшими.

3.2. Модели стационарных временных рядов

Стационарный временной ряд — это временной ряд (3.2), вероятностные свойства которого *со временем не изменяются*. Более точно стационарным называют временной ряд, у которого математическое ожидание, дисперсия и ковариации наблюдений y_t не зависят от времени:

$$M(y_t) = const,\ D(y_t) = \sigma^2 = const,\ cov(y_t, y_{t-k}) = c_k.$$

Стационарный временной ряд получается из временного ряда (3.3) удалением неслучайной составляющей $f_t + \alpha_t + \beta_t$, за исключением, возможно, постоянного слагаемого.

3.2.1. Авторегрессионные модели

При выборе структуры временного ряда часто применяют регрессионные модели с *лаговыми* переменными, относящимися к некоторым *предыдущим* моментам времени. В их числе – *авторегрессионные* модели p-го порядка ($p \in \mathbb{N}$)

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_p y_{t-p} + \varepsilon_t \quad (t \in \overline{p+1, n}), \tag{3.9}$$

где $\varepsilon_t \sim N(0, \sigma^2)$. Модели (3.9) кратко обозначают AR(p) от англ. «autoregressive» (авторегрессионный).

Рассмотрим подробнее простейшую модель этого класса – модель AR(1):

$$y_t = b_0 + b_1 y_{t-1} + \varepsilon_t. \tag{3.10}$$

Условием стационарности временного ряда (3.10) является неравенство

$$|b_1| < 1. \tag{3.11}$$

При $|b_1| > 1$ модель (3.10) в реальных эконометрических задачах не встречается [14, с. 279].

Будем исследовать модель (3.10) как *модель парной линейной регрессии с дискретным регрессором*

$$x_t = y_{t-1}.$$

При этом $t \in \overline{2, n}$, так что число наблюдений в модели (3.10) равно $n-1$. Выборка (1.5) принимает специфический вид

$$(y_1, y_2), (y_2, y_3), \ldots, (y_{n-1}, y_n). \tag{3.12}$$

Точечные оценки $\widehat{b}_1, \widehat{b}_0$ коэффициентов b_1, b_0 определяются с помощью МНК так же, как и в подп. 1.2.1, с тем лишь исключением, что для вычисления средних выборочных значений вместо формул (1.19) надо применять формулы

$$\begin{gathered} \overline{x} = \frac{\sum\limits_{t=2}^{n} x_t}{n-1} = \frac{\sum\limits_{t=2}^{n} y_{t-1}}{n-1}, \ \overline{x^2} = \frac{\sum\limits_{t=2}^{n} y_{t-1}^2}{n-1}, \\ \overline{y} = \frac{\sum\limits_{t=2}^{n} y_t}{n-1}, \ \overline{xy} = \frac{\sum\limits_{t=2}^{n} y_{t-1} y_t}{n-1}, \ \overline{y^2} = \frac{\sum\limits_{t=2}^{n} y_t^2}{n-1}. \end{gathered} \tag{3.13}$$

Исследование авторегрессионных моделей $AR(p)$ существенно зависит от параметра p и опирается на аппарат теории случайных функций [1, с. 822–837] .

Пример 3.2[2]. Временной ряд определяет динамику курса акций некоторой компании (табл. 3.3).

Таблица 3.3

t	1	2	3	4	5	6	7	8	9	10	11
y_t	971	1166	1044	907	957	727	752	1019	972	815	823
t	12	13	14	15	16	17	18	19	20	21	22
y_t	1112	1386	1428	1364	1241	1145	1351	1325	1226	1189	1213

Требуется построить $AR(1)$ модель этого временного ряда, оценить ее качество на уровне значимости $\alpha = 0,05$; сравнить на плоскости (t, y) графики исходного временного ряда и сглаженного по модели (3.10) временного ряда $\{\widehat{y}_t\}$.

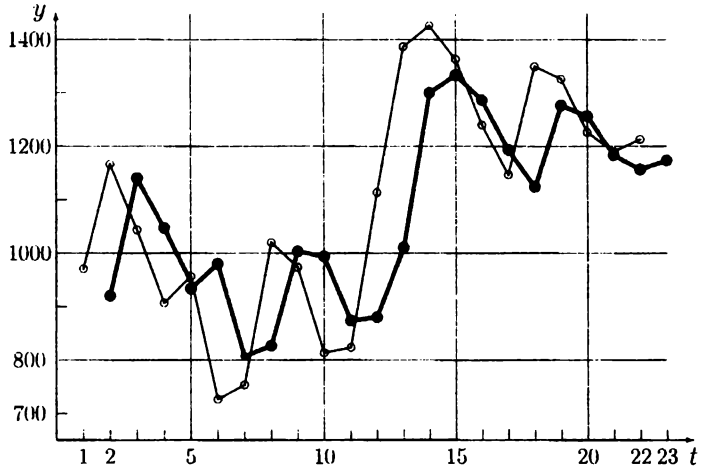


Рис. 3.3. Авторегрессионная модель. Пример 3.2

Решение. Построим график исходного временного ряда (рис. 3.3), отмечая кружками точки корреляционного поля. Динамика временного ряда сложна, ее нельзя качественно представить трендовой моделью.

[2]См. работу Н.Ш. Кремера и Б.А. Путко [12, с. 147].

Таблица 3.4

t	2	3	4	5	...	20	21	22
y_{t-1}	971	1166	1044	907	...	1325	1226	1189
y_t	1166	1044	907	957	...	1226	1189	1213

Применим авторегрессионную модель первого порядка (3.10). Это модель парной линейной регрессии с регрессором y_{t-1}. Здесь $t \in \overline{2,22}$. Для регрессионного уравнения (3.10) выборка записана в табл. 3.4. Применив МНК, находим точечные оценки коэффициентов регрессии

$$\widehat{b}_0 = 261,593;\ \widehat{b}_1 = 0,752.$$

Таким образом, оцененное уравнение регрессии имеет вид

$$\widehat{y}_t = 261,593 + 0,752y_{t-1}.$$

Подставляя в эту формулу значения $t = 2, 3, \ldots, 22$, заполняем табл. 3.5.

Таблица 3.5

t	2	3	4	5	6	7	8
$\widehat{y}_t$	992,15	1139	1047	944	981,62	808,57	827,38
t	9	10	11	12	13	14	15
$\widehat{y}_t$	1028	992,3	874,78	881	1098	1304	1336
t	16	17	18	19	20	21	22
$\widehat{y}_t$	1288	1195	1123	1278	1258	1184	1156

Появился новый временной ряд, определенный моделью (3.10). Его график на рис. 3.3 отмечен более жирной ломаной. Сравнение обоих графиков показывает, что модель достаточно точно (с минимальным запаздыванием) отслеживает изучаемый процесс. Интересно, что оцененный ряд дает прогноз на один шаг – значение $\widehat{y}_{23}$, равное 1174.

Изучим качество модели, используя коэффициент детерминации. Находим $R^2 = 0,565$. Вычисляем

$$F_{\text{набл}} \overset{(1.82)}{=} \frac{R^2}{1-R^2} \cdot (n-2) = \frac{0,565}{0,435} \cdot 19 = 24,678;$$

$$F_{\text{крит}} = F_{1-\alpha}(1, n-2) = F_{0,95}(1, 19) = 4,38 < F_{\text{набл}}.$$

Таким образом, уравнение (3.10) на уровне значимости $\alpha = 0,05$ соответствует наблюдениям.

3.2.2. Модели распределенных лагов

Рассмотрим теперь такие модели, где запаздывание (лагирование) распространяется на *независимую переменную* x :

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_p x_{t-p} + \varepsilon_t \quad (t \in \overline{p+1, n}). \tag{3.14}$$

В модели (3.14) одновременно участвуют два стационарных временных ряда

$$\{x_t\},\ \{y_t\} \qquad (t \in \overline{1, n}),$$

причем изучается зависимость переменной y от переменной x.

Модель (3.14) применяют в тех ситуациях, когда переменная y с запаздыванием реагирует на изменения регрессора x. Ее называют моделью *распределенных лагов* порядка p или, кратко, моделью типа $DL(p)$ (от англ. «distributed lags»).

Выясним смысл коэффициентов b_i. Влияние переменной x на переменную y сохраняется в течение p шагов.

Коэффициент b_0 называют *краткосрочным мультипликатором*. Он характеризует среднее изменение переменной y в момент t, если в этот момент аргумент x увеличился на 1, а его лаговые значения не учитываются.

Величина

$$b = b_0 + b_1 + \ldots + b_p$$

называется *долгосрочным мультипликатором*. Она определяет суммарное изменение переменной y к моменту $t+p$, если в момент t аргумент x увеличился на 1. Вклад отдельного i-го лага в это суммарное изменение равен $w_i = b_i/b$, при этом $\sum\limits_{i=0}^{p} w_i = 1$. Конечная последовательность

$$w_0, w_1, \ldots . w_p$$

называется *распределением лагов*. Скорость реакции переменной y на изменение регрессора x можно охарактеризовать с помощью *среднего лага*

$$l_{\text{ср}} = \sum_{i=0}^{p} i\, w_i. \tag{3.15}$$

Малые значения среднего лага означают быструю реакцию y на изменение x, и наоборот.

В принципе, модель (3.14) можно исследовать как модель множественной регрессии с регрессорами

$$x_t,\ x_{t-1},\ \ldots,\ x_{t-p}, \tag{3.16}$$

их число - $(p+1)$. Однако на практике такой прямой подход часто осложняется мультиколлинеарностью модели, поскольку наблюдения (3.16) склонны коррелировать между собой. Поэтому обычно применяют специальные методы, *ослабляющие мультиколлинеарность и уменьшающие число регрессоров.* Основные из них – метод полиномиальных лагов и метод геометрических лагов.

3.2.3. Метод полиномиальных лагов

Данный метод принадлежит эконометристу Ширли Алмон. Она предложила выбирать коэффициенты b_i как *значения в точке i многочлена фиксированной степени* $m < p$:

$$b_i = \gamma_0 + \gamma_1 i + \ldots + \gamma_m i^m. \tag{3.17}$$

Подставляя формулы (3.17) в (3.14), получим модель множественной регрессии:

$$y_t = a + \gamma_0 z_{t0} + \gamma_1 z_{t1} + \ldots + \gamma_m z_{tm} + \varepsilon_t. \tag{3.18}$$

Новые дискретные регрессоры

$$z_{t0}, z_{t1}, \ldots, z_{tm}$$

являются линейными комбинациями переменных (3.16). Коэффициенты

$$\gamma_0,\ \gamma_1,\ \ldots,\ \gamma_m$$

модели (3.18) оценивают с помощью обычного МНК. Затем по формулам (3.17) находят оценки *исходных коэффициентов* b_i. Обычно выбирают $m = 2$ или $m = 3$.

Найдем формулы перехода от модели (3.14) к модели (3.18) в случае

$$p = 3,\ m = 2.$$

Тогда исходная модель принимает вид

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2, x_{t-2} + b_3 x_{t-3} + \varepsilon_t \quad (t \in \overline{4,n}), \tag{3.19}$$

а ее коэффициенты, согласно формулам (3.17), записываются так:

$$\begin{aligned} b_0 &= \gamma_0, \\ b_1 &= \gamma_0 + \gamma_1 + \gamma_2, \\ b_2 &= \gamma_0 + 2\gamma_1 + 4\gamma_2, \\ b_3 &= \gamma_0 + 3\gamma_1 + 9\gamma_2. \end{aligned} \tag{3.20}$$

Преобразуем

$$b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + b_3 x_{t-3} \overset{(3.20)}{=} \gamma_0 x_t + (\gamma_0 + \gamma_1 + \gamma_2) x_{t-1} + (\gamma_0 +$$

$$+2\gamma_1 + 4\gamma_2) x_{t-2} + (\gamma_0 + 3\gamma_1 + 9\gamma_2) x_{t-3} = \gamma_0 (x_t + x_{t-1} + x_{t-2} + x_{t-3}) +$$

$$+\gamma_1 (x_{t-1} + 2x_{t-2} + 3x_{t-3}) + \gamma_2 (x_{t-1} + 4x_{t-2} + 9x_{t-3}) = \gamma_0 z_{t0} + \gamma_1 z_{t1} + \gamma_2 z_{t2}.$$

Здесь

$$\begin{aligned} z_{t0} &= x_t + x_{t-1} + x_{t-2} + x_{t-3}, \\ z_{t1} &= x_{t-1} + 2x_{t-2} + 3x_{t-3}, \\ z_{t2} &= x_{t-1} + 4x_{t-2} + 9x_{t-3}. \end{aligned} \tag{3.21}$$

Таким образом, пришли к множественной модели с тремя регрессорами

$$y_t = a + \gamma_0 z_{t0} + \gamma_1 z_{t1} + \gamma_2 z_{t2} + \varepsilon_t \quad (t \in \overline{4,n}). \tag{3.22}$$

Пример 3.3[3]. Известна динамика изменения в течение 13 лет объема валового внутреннего продукта y некоторой страны в зависимости от инвестиций x в ее экономику (табл. 3.6).

Таблица 3.6

t	1	2	3	4	5	6	7	8	9	10	11	12	13
x_t	30	29	29	32	34	37	41	44	42	44	46	43	48
y_t	193	197	202	213	222	234	247	262	269	280	287	287	296

Требуется:

1) Применить метод полиномиальных лагов при $p = 3$, $m = 2$.

[3]См. работу А.И. Новикова [16, с. 82].

2) Вычислить выборочные значения мультипликаторов и среднего лага.

Решение:

1) Найдем по формулам (3.21) значения новых переменных z_{t0}, z_{t1}, z_{t2} $(4 \leqslant t \leqslant 13)$. Например,

$$z_{40} = x_4 + x_3 + x_2 + x_1 = = 32 + 29 + 29 + 30 = 120,$$

$$z_{41} = x_3 + 2x_2 + 3x_1 = 177,\ z_{42} = x_3 + 4x_2 + + 9x_1 = 415.$$

Теперь сможем заполнить табл. 3.7. определяющую выборку множественной регрессии (3.22).

Таблица 3.7

t	y_t	z_{t0}	z_{t1}	z_{t2}
4	213	120	177	415
5	222	124	177	409
6	234	132	185	423
7	247	144	201	461
8	252	156	217	495
9	269	164	237	541
10	280	171	253	587
11	287	176	260	608
12	287	175	260	600
13	296	181	267	623

С помощью обычного МНК находим точечные оценки параметров множественной регрессии. Введем регрессионную матрицу $\mathbf{Z}$ и вектор $\boldsymbol{\gamma}$ коэффициентов регрессии

$$\mathbf{Z} = \begin{pmatrix} 1 & 120 & 177 & 415 \\ 1 & 124 & 177 & 409 \\ \dots & \dots & \dots & \dots \\ 1 & 181 & 267 & 623 \end{pmatrix},\ \boldsymbol{\gamma} = \begin{pmatrix} a \\ \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

Тогда $\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = (57,886;\ 1,658;\ -0,61;\ 0,159)^T$. Таким образом,

$$\widehat{a} = 57,886,\ \widehat{\gamma}_0 = 1,658,\ \widehat{\gamma}_1 = -0,61,\ \widehat{\gamma}_2 = 0,159.$$

Теперь по формулам (3.20), где символы всех величин снабжены «крышками», определяем точечные оценки коэффициентов b_0, b_1, b_2, b_3 исходной модели (3.19):

$$\widehat{b}_0 = \widehat{\gamma}_0 = 1,658;\ \widehat{b}_1 = \widehat{\gamma}_0 + \widehat{\gamma}_1 + \widehat{\gamma}_2 = 1,207;\ \widehat{b}_2 = 1,074;\ \widehat{b}_3 = 1,259.$$

Таким образом, оцененная по выборке модель распределенных лагов имеет вид

$$\widehat{y}_t = 57,886 + 1,658x_t + 1,207x_{t-1} + 1,074x_{t-2} + 1,259x_{t-3}.$$

2) Находим выборочные оценки краткосрочного и долгосрочного мультипликаторов:

$$\widehat{b}_0 = 1,658,\ \widehat{b} = 1,658 + 1,207 + 1,074 + 1,259 = 5,198$$

соответственно. Поэтому увеличение инвестиций на 1 у.е. приведет к росту ВВП в среднем на 1,658 у.е. в текущем году и на 5,198 у.е. через 3 года.

Найдем распределение лагов: $\widehat{w}_0 = \widehat{b}_0/\widehat{b} = 0,319;\ \widehat{w}_1 = 0,232;\ \widehat{w}_2 = 0,207,\ \widehat{w}_3 = 0,242$. Вычислим средний лаг $\widehat{l}_{\text{ср}}$ по формуле (3.15):

$$\widehat{l}_{\text{ср}} = 0,232 + 2 \cdot 0,207 + \ + 3 \cdot 0,242 = 1,372.$$

Следовательно, увеличение инвестиций в экономику страны приведет к увеличению ВВП в среднем через 1,372 года.

3.2.4. Метод геометрических лагов

Если порядок p модели (3.14) велик, то удобно считать его бесконечно большим и рассматривать модель

$$y_t = a + b_0x_t + b_1x_{t-1} + \ldots + b_px_{t-p} + \ldots + \varepsilon_t.$$

Здесь влияние y на x продолжается бесконечно.

В методе *геометрических лагов* предполагается, что с увеличением величины лага p коэффициенты b_p изменяются по закону бесконечно убывающей геометрической прогрессии:

$$b_p = s\lambda^p \quad (0 < \lambda < 1).$$

Знаменатель прогрессии λ характеризует скорость убывания параметров $|b_p|$; при малых λ она больше.

Модель геометрических лагов записывается в виде

$$y_t = a + sx_t + s\lambda x_{t-1} + \ldots + s\lambda^p x_{t-p} + \ldots + \varepsilon_t. \qquad (3.23)$$

Краткосрочный мультипликатор b_0 модели (3.23) – коэффициент s : $b_0 = s$. Долгосрочный мультипликатор определяется следующим образом:

$$b = s\sum_{p=0}^{\infty} \lambda^p = \frac{s}{1-\lambda}.$$

Здесь применена формула суммы всех членов бесконечно убывающей геометрической прогрессии

$$\sum_{p=0}^{\infty} \lambda^p = \frac{1}{1-\lambda}.$$

Модель (3.23) содержит три параметра s, λ, a. Модель *нелинейна*, поскольку в нее входят *произведения* $s\lambda^p$.

Рассмотрим один из способов исследования модели геометрических лагов (3.23):

1. Перебираются с некоторым шагом значения параметра λ из интервала $(0, 1)$.

2. Для каждого значения λ определяется переменная

$$z_t = x_t + \lambda x_{t-1} + \ldots + \lambda^p x_{t-p}$$

с таким значением p, что дальнейшие лаговые значения x_{t-k}, $k > p$ не оказывают существенного воздействия на z_t. Это возможно потому, что в силу неравенств $0 < \lambda < 1 \quad \lim\limits_{p\to\infty} \lambda^p = 0$ и числа λ^p при достаточно больших p сколь угодно близки к нулю. Значит, можно перейти от модели (3.23) к модели с *конечным числом лагов*

$$y_t = a + sx_t + s\lambda x_{t-1} + \ldots + s\lambda^p x_{t-p} + \varepsilon_t.$$

Это уравнение можно записать как уравнение парной линейной регрессии

$$y_t = a + sz_t + \varepsilon_t. \qquad (3.24)$$

Коэффициенты a, s уравнения (3.24) оцениваются с помощью обычного МНК.

3. Выбирается то значение параметра λ, которое обеспечивает *наибольший* коэффициент детерминации R^2 при оценке уравнения (3.24).

Замечание 3.1. Большое практическое значение имеют модели временных рядов, в которых запаздывание распространяется и на регрессор, и на зависимую переменную:

$$y_t = b_0 + b_1 y_{t-1} + \ldots + b_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_q x_{t-q} + \varepsilon_t.$$

Такие модели называют авторегрессионными моделями с распределенными лагами порядков p, q или моделями $ADL(p, q)$. Названные модели весьма сложны, достаточно полное их исследование возможно лишь при малых p, q [6, с. 303–320].

Замечание 3.2. Мы рассмотрели некоторые методы исследования стационарных временных рядов. При изучении *нестационарных временных рядов* в различных случаях применяются разные подходы. Укажем некоторые из них:

1) Выделение и удаление неслучайных компонент временного ряда.

2) Переход от исходного временного ряда к временному ряду его *разностей* некоторого порядка (см. п. 3.1).

3) Методы *коинтеграции* – объединения нескольких нестационарных временных рядов в их стационарную линейную комбинацию [1, с. 866 - 867; 14, с. 283 284].

3.3. Адаптивные модели прогнозирования временных рядов

Прогнозированием временного ряда называют предсказание будущих значений показателей изучаемого экономического процесса. В основе методов прогнозирования лежит предположение о том, что основные факторы и тенденции, имевшие место в прошлом, сохраняются и в будущем. При этом наиболее важными зачастую оказываются *последние уровни* ряда, а более ранняя информация о процессе имеет меньший вес, устаревает.

Мы будем рассматривать одномерные временные ряды. Адаптивные модели прогнозирования – это модели, непрерывно подстраивающиеся под динамику процесса. В отличие от трендовых моделей теперь различным уровням ряда присваиваются разные веса.

Общая схема адаптивных моделей такова:

1) По нескольким первым уровням ряда находятся начальные значения параметров модели.

2) Начиная с первого уровня строится прогноз на один шаг, сравнивается с истинным уровнем временного ряда, а ошибка прогноза учитывается при корректировке параметров модели и т.д. Поэтому модель на каждом шаге обновляется.

3) Последние значения параметров модели позволяют дать краткосрочные и среднесрочные прогнозы.

Рассмотрим подробнее часто применяемую адаптивную модель – *модель экспоненциального сглаживания Брауна.* Это такая модель взвешенного скользящего среднего, что веса прошлых наблюдений экспоненциально уменьшаются по мере их удаления от последнего используемого наблюдения.

Существуют различные модификации метода Брауна, отражающие развитие процесса с линейной или квадратичной тенденцией.

Рассмотрим *линейный метод Брауна.* В основе метода лежит формула прогноза на k шагов вперед исходя из любого момента времени t, $t \in \overline{0,n}$:

$$\widehat{y}_{t,k} = A_0(t) + A_1(t)\,k. \tag{3.25}$$

Конкретизируем для метода Брауна общую схему адаптивных моделей:

1) По первым m уровням временного ряда с помощью обычного МНК оцениваются коэффициенты линейного тренда, которые и выбираются в качестве начальных коэффициентов $A_0(0), A_1(0)$ модели Брауна.

2) Остальные коэффициенты

$$A_0(t),\ A_1(t) \quad (t \in \overline{1,n})$$

последовательно определяются по однотипным формулам

$$A_0(t) = A_0(t-1) + A_1(t-1) + (1-\lambda)e_t;\ A_1(t) = A_1(t-1) + \lambda e_t. \tag{3.26}$$

Здесь величина

$$e_t = y_t - \widehat{y}_{t-1,1} \tag{3.27}$$

является ошибкой прогноза, расхождением между наблюдаемым значением y_t переменной y в момент времени t и прогнозируемым ее значением $\widehat{y}_{t-1,1}$, полученным по информации о модели в предыдущий момент

$t-1$. Согласно формуле (3.25)

$$\widehat{y}_{t-1.1} = A_0(t-1) + A_1(t-1).$$

Коэффициенты $A_0(t), A_1(t)$ корректируются на каждом шаге. В формулы (3.26) входит коэффициент адаптации λ $(0 < \lambda < 1)$.

3) После нахождения последних коэффициентов $A_0(n), A_1(n)$ строится прогноз на k шагов по формуле

$$\widehat{y}_{n,k} = A_0(n) + A_1(n)\,k. \tag{3.28}$$

Формула (3.28) даст точечную оценку прогноза. Используя подход, описанный в подп. 1.3.6, можно найти и интервальные оценки.

Пример 3.4. Применим адаптивную линейную модель Брауна для прогнозирования курса евро относительно рубля. Известны последовательные наблюдения за курсом евро в ноябре 2006 г. (табл. 3.8)[4].

Таблица 3.8

t	1	2	3	4	5	6	7
y_t	33,99	34,08	34,08	34,11	34,09	34,1	34,11
t	8	9	10	11	12	13	14
y_t	34,24	34,23	34,18	34,17	34,17	34,12	34,19
t	15	16	17	18	19	20	21
y_t	34,17	34,25	34,36	34,4	34,6	34,63	34,68

Требуется:

А. Выбирая $m = 5, \lambda = 0,16$, по первым 19 наблюдениям применить линейный метод Брауна. Найти прогнозы на 1 шаг и на 2 шага вперед, сравнить прогнозные и реальные значения.

Б. Построить на плоскости (t, y) графики исходного временного ряда и временного ряда метода Брауна, включая прогнозы.

Решение:

А. В нашем случае $n = 19$.

1) Строя по первым 5 наблюдениям линейный тренд

$$\widehat{y}_t = A_0 + A_1 t$$

[4]См.: Рос. газ. 2006. 1– 30 нояб.

и оценивая обычным МНК его коэффициенты, находим $\widehat{A}_0 = 34,001$; $\widehat{A}_1 = 0,023$. Поэтому

$$A_0(0) = \widehat{A}_0 = 34,001;\ A_1(0) = \widehat{A}_1 = 0,023.$$

2) Формулы (3.26) преобразуются к виду

$$A_0(t) = A_0(t-1) + A_1(t-1) + 0,84 \cdot e_t,\ A_1(t) = A_1(t-1) + 0,16 \cdot e_t.$$

Заполняем табл. 3.9. Из нее находим $A_0(19) = 34,573$; $A_1(19) = 0,064$.

Таблица 3.9

t	y_t	$A_0(t)$	$A_1(t)$	$\widehat{y}_{t-1,1}$	e_t
0		34,001	0,023		–
1	33,99	33,995	0,018	34,024	– 0,034
2	34,08	34,069	0,028	34,013	0,067
3	34,08	34,083	0,025	34,098	0,018
4	34,11	34,11	0,026	34,108	0,002
5	34,09	34,098	0,018	34,135	– 0,045
6	34,1	34,103	0,016	34,116	– 0,016
7	34,11	34,111	0,015	34,118	– 0,008
8	34,24	34,222	0,033	34,126	0,114
9	34,23	34,234	0,029	34,255	– 0,025
10	34,18	34,193	0,016	34,263	0,083
11	34,17	34,176	0,009	34,209	– 0,039
12	34,17	34,173	0,007	34,186	– 0,016
13	34,12	34,13	0,003	34,179	0,059
14	34,19	34,18	0,008	34,127	0,063
15	34,17	34,173	0,005	34,187	– 0,017
16	34,25	34,238	0,016	34,178	0,072
17	34,36	34,343	0,033	34,255	0,105
18	34,4	34,396	0,037	34,376	0,024
19	34,6	34,573	0,064	34,433	0,167

3) Запишем формулу прогноза на k шагов вперед:

$$\widehat{y}_{19,k} \overset{(3.28)}{=} 34,573 + 0,064k.$$

Вычисляем по этой формуле требуемые прогнозы, округляя их с выбранной точностью 0,01:

$$\widehat{y}_{20} = \widehat{y}_{19,1} = 34,573 + 0,064 = 34,64;$$

$$\widehat{y}_{21} = \widehat{y}_{19,2} = 34,573 + 0,128 = 34,70.$$

Ошибка первого и второго прогнозов равна 0,01 и 0,02 соответственно.

Б. Графики исходного временного ряда и выровненного по методу Брауна временного ряда построены на рис. 3.4. Второй график обозначен более жирной линией.

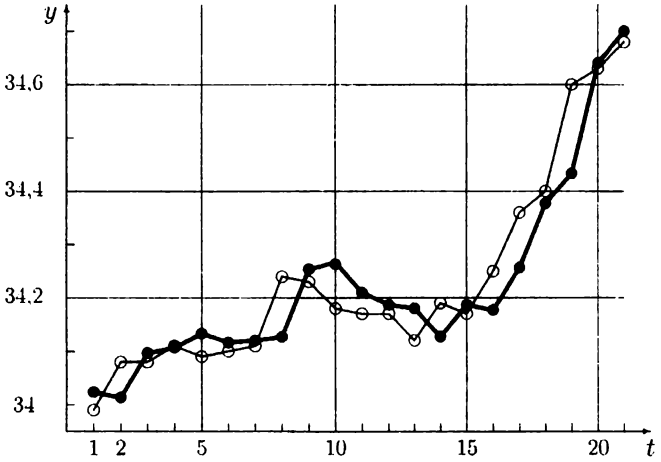


Рис. 3.4. Линейная модель прогнозирования Брауна. Пример 3.4

Замечание 3.3. Метод Брауна — простейший из адаптивных методов прогнозирования. В эконометрической практике применяются и другие модели прогнозирования: метод Хольта с двумя коэффициентами адаптации; методы Хольта–Уинтерса и Тейла–Вейджа, учитывающие сезонные колебания [1, с. 901–903].

Контрольные задания

1. Укажите компоненты одномерного временного ряда.

2. Запишите формулы нахождения МНК-оценок параметров модели квадратичного тренда

$$y_t = b_0 + b_1 t + b_2 t^2 + \varepsilon_t.$$

3. Запишите формулы преобразования модели распределенных лагов (3.14) по методу полиномиальных лагов в случае $p = 4, m = 2$.

4. Примените линейную модель Брауна для прогнозирования курса доллара относительно рубля по следующим последовательным наблюдениям в ноябре 2006 г. (таблица)[5].

t	1	2	3	4	5	6	7
y_t	26,78	26,72	26,73	26,7	26,72	26,7	26,7
t	8	9	10	11	12	13	14
y_t	26,62	26,62	26,65	26,65	26,66	26,69	26,64
t	15	16	17	18	19	20	21
y_t	26,65	26,61	26,56	26,52	26,37	26,35	26,31

[5]См.: Рос. газ. 2006. 1– 30 нояб.

4. Линейные системы одновременных уравнений

До сих пор мы изучали регрессионные модели, состоящие из одного уравнения. Такими уравнениями описываются зависимости между одной зависимой переменной и независимыми переменными. Однако сложные экономические объекты, особенно в макроэкономике, моделируются системами нескольких взаимосвязанных уравнений.

Уравнения этих систем нельзя рассматривать изолированно, поскольку одна и та же переменная в одних уравнениях является независимой, а в других — зависимой. Она одновременно играет разные роли, поэтому в эконометрике принят термин *системы одновременных уравнений* (СОУ).

4.1. Классификация переменных в системах одновременных уравнений

Мы ограничимся изучением *линейных* систем одновременных уравнений[1] . Будем считать также, что статистические данные об экономическом объекте являются временны́ми. Поэтому рассматриваемые СОУ являются обобщением временны́х рядов, при этом всегда $t \in \overline{1,n}$.

Проведем классификацию переменных СОУ (рисунок). Переменные делятся на *эндогенные* и *предопределенные*.

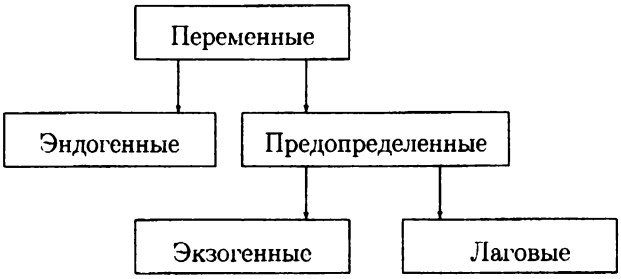
Значения *эндогенных* переменных определяются *внутри модели*. Эндогенные переменные являются зависимыми переменными, однако в некоторые уравнения системы формально входят как независимые переменные.

---

[1]Теория нелинейных систем одновременных уравнений в достаточной степени не разработана и в учебной литературе по эконометрике обычно не рассматривается.

Значения *экзогенных* переменных определяются *вне модели*. Эти переменные влияют на эндогенные переменные, но не зависят от них.

В систему одновременных уравнений могут входить *лаговые переменные* – значения переменных в предыдущих измерениях. *Предопределенные* переменные *включают экзогенные и лаговые переменные*. Они играют роль аргументов, объясняющих переменных.



Переменные в системах одновременных уравнений

Система одновременных уравнений обязательно содержит *регрессионные (поведенческие) уравнения*. Она может включать также *тождества*, связывающие некоторые переменные и не содержащие случайных компонент.

4.2. Примеры систем одновременных уравнений

Построим некоторые простые, практически важные примеры систем одновременных уравнений.

Пример 4.1. (модель спроса-предложения). Рассмотрим классическую систему, используемую при моделировании спроса-предложения в рыночной экономике.

Вводятся переменные: q^D – спрос на определенный товар, q^S – предложение товара, p – цена товара, y – доход потребителей. Система содержит два регрессионных уравнения и одно тождество

$$\begin{cases} q_t^D = a_1 p_t + b_1 + b_2 y_t + \varepsilon_{t1}, \\ q_t^S = a_2 p_t + b_3 + \varepsilon_{t2}, \\ q_t^D = q_t^S. \end{cases} \tag{4.1}$$

Первое уравнение определяет функцию спроса, второе – функцию предложения, третье уравнение задает условие равновесия спроса и предложения. В рассматриваемой модели q_t^D, q_t^S, p_t – эндогенные переменные, y_t – экзогенная переменная.

Существуют разновидности моделей спроса-предложения, учитывающие различные обстоятельства. Например, уравнение спроса может дополнительно учитывать инерцию цен:

$$q_t^D = a_1 p_t + b_0 + b_1 p_{t-1} + b_2 y_t + \varepsilon_{t1}.$$

Здесь появилась лаговая эндогенная переменная p_{t-1}. Таким образом, предопределенные переменные новой модели y_t, p_{t-1}.

Пример 4.2. (кейнсианская модель формирования доходов). Рассмотрим макроэкономическую модель закрытой экономики (без государственных расходов). Пусть y – совокупный выпуск, c – объем потребления, d – объем инвестиций. Система состоит из одного регрессионного уравнения, определяющего функцию потребления, и макроэкономического тождества:

$$\begin{cases} c_t = a y_t + b + \varepsilon_t \quad (0 < a < 1), \\ y_t = c_t + d_t. \end{cases} \tag{4.2}$$

Здесь y_t, c_t – эндогенные переменные, d_t – экзогенная переменная.

На последнем примере укажем основные подходы к исследованию систем одновременных уравнений.

Исходная система (4.2) называется *структурной формой* рассматриваемой эконометрической модели. Коэффициенты a, b называют *структурными коэффициентами* модели. Одна из основных задач исследования модели – оценка структурных коэффициентов. Для этого нужно найти *приведенную форму* модели, т.е. выразить из системы (4.2) эндогенные переменные y_t, c_t через экзогенную переменную d_t и случайное слагаемое ε_t. Подставим второе уравнение системы в первое и преобразуем:

$$c_t = b + a c_t + a d_t + \varepsilon_t, \; c_t = \frac{b}{1-a} + \frac{a}{1-a} d_t + \frac{\varepsilon_t}{1-a};$$

$$y_t = \frac{b}{1-a} + \frac{1}{1-a} d_t + \frac{\varepsilon_t}{1-a}.$$

Полагая

$$\gamma_1 = \frac{a}{1-a}\,,\ \gamma_2 = \frac{b}{1-a}\,; \qquad (4.3)$$

$$u_t = \frac{\varepsilon_t}{1-a}\,,$$

приходим к следующей приведенной форме модели:

$$\begin{cases} c_t = \gamma_1 d_t + \gamma_2 + u_t, \\ y_t = (1+\gamma_1) d_t + \gamma_2 + u_t. \end{cases} \qquad (4.4)$$

Здесь γ_1, γ_2 – приведенные коэффициенты модели.

Необходимость рассмотрения приведенной формы (4.4) объясняется тем, что применение МНК к первому уравнению системы (4.2) неэффективно. Действительно, переменная y_t, будучи регрессором для этого уравнения, является случайной величиной, зависящей от случайных величин c_t, ε_t. Последнее вытекает из второго уравнения системы (4.2). Поэтому не выполнено 1^0 условие Гаусса–Маркова и МНК даст смещенные и несостоятельные оценки структурных коэффициентов a, b.

В таких случаях применяют усложненный вариант МНК, называемый *косвенным методом наименьших квадратов*. Он основан на совместном использовании обеих форм модели – структурной и приведенной.

Существенно, что из равенств (4.3), выражающих приведенные уравнения через структурные, можно однозначно определить, наоборот, структурные коэффициенты через приведенные:

$$a = \frac{\gamma_1}{1+\gamma_1}\,,\ b = \frac{\gamma_2}{1+\gamma_1}\,. \qquad (4.5)$$

В таких случаях говорят, что исходная, структурная система *идентифицируема*.

В то же время уравнения приведенной формы (4.4) модели удовлетворяют условиям Гаусса–Маркова. Это обусловлено тем, что их регрессор d_t – независимая, экзогенная переменная, не коррелирующая с ε_t, а значит, и с u_t. Поэтому, применяя к одному из уравнений системы (4.4) обычный МНК, находим оценки структурных коэффициентов, обладающие требуемыми статистическими свойствами. После этого, воспользовавшись формулами (4.5), определяем несмещенные и состоятельные оценки структурных коэффициентов.

Применим косвенный МНК к конкретной модели вида (4.2).

Пример 4.3[2]. Для некоторой страны имеются следующие данные за десятилетие (табл. 4.1).

Таблица 4.1

y_t	200	218	230	200	210	230	250	230	220	240
c_t	190	198	200	180	200	210	220	210	205	210

Требуется сравнить два способа оценки структурных коэффициентов a, b :

1) с помощью косвенного МНК;

2) с помощью обычного МНК (используя лишь исходную модель (4.2)).

Решение:

1) Так как

$$d_t \overset{(4.2)}{=} y_t - c_t,$$

то по исходной таблице определяем выборочные значения эндогенной переменной d_t (табл. 4.2).

Таблица 4.2

d_t	10	20	30	20	10	20	30	20	15	30

Найдем МНК-оценки коэффициентов первого уравнения приведенной модели (4.4). По формулам парной линейной регрессии (регрессор d_t) вычисляем

$$\widehat{\gamma}_1 = 0,696, \ \widehat{\gamma}_2 = 188,038.$$

Применяя формулы (4.5), где всюду добавлены «крышки», находим оценки структурных коэффициентов:

$$\widehat{a} = \frac{0,696}{1+0,696} = 0,41, \quad \widehat{b} = \frac{188,038}{1+0,696} = 110,87.$$

Таким образом,

$$\widehat{a} = 0,41, \quad \widehat{b} = 110,87. \tag{4.6}$$

[2]См. работу А.И. Новикова [16, с. 99].

Напомним, что эти оценки несмещенные и состоятельные.

2) Оценим структурные коэффициенты непосредственно – из первого уравнения исходной модели (4.2). Вновь применяя процедуру парной линейной регрессии (теперь регрессор – y_t), находим

$$\widehat{a} = 0,635, \quad \widehat{b} = 60,882. \tag{4.7}$$

Отличия оценок (4.6), (4.7) значительные.

4.3. Структурная и приведенная формы систем одновременных уравнений

Укажем общий вид линейной системы одновременных уравнений, являющейся эконометрической моделью рассматриваемого сложного экономического объекта. Если в систему входят *тождества*, то они позволяют *исключить* некоторые переменные и уравнения. Будем предполагать, что это уже произошло. Пусть система содержит эндогенные переменные

$$y_1, y_2, \ldots, y_k \quad (k \geqslant 2)$$

и предопределенные переменные

$$x_1, x_2, \ldots, x_m.$$

В число предопределенных переменных включим константу $x_1 = 1$; коэффициенты при ней будут *свободными членами* СОУ.

Обозначим выборочные значения переменных в моменты наблюдений t $(t \in \overline{1,n})$ так:

$$y_{t1}, y_{t2}, \ldots, y_{tk};\ x_{t1}, x_{t2}, \ldots, x_{tm}$$

$(x_{t1} = 1)$. Систему одновременных уравнений запишем в виде

$$\begin{cases} y_{t1} + a_{12}y_{t2} + \ldots + a_{1k}y_{tk} + b_{11} + b_{12}x_{t2} + \ldots + b_{1m}x_{tm} = \varepsilon_1, \\ a_{21}y_{t1} + y_{t2} + \ldots + a_{2k}y_{tk} + b_{21} + b_{22}x_{t2} + \ldots + b_{2m}x_{tm} = \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{k1}y_{t1} + a_{k2}y_{t2} + \ldots + y_{tk} + b_{k1} + b_{k2}x_{t2} + \ldots + b_{km}x_{tm} = \varepsilon_k. \end{cases} \tag{4.8}$$

Число уравнений в системе равно числу эндогенных переменных.

Систему (4.8) называют структурной формой эконометрической модели.

Некоторые из коэффициентов a_{ij}, b_{ij} могут быть нулями (обычно нулей достаточно много). Однако в каждое i-е уравнение системы (4.8) обязательно входит эндогенная переменная y_i. Поэтому систему можно представить в другой форме:

$$\begin{cases} y_{t1} = -a_{12}y_{t2} - \ldots - a_{1k}y_{tk} - b_{11} - b_{12}x_{t2} - \ldots - b_{1m}x_{tm} + \varepsilon_1, \\ y_{t2} = -a_{21}y_{t1} - \ldots - a_{2k}y_{tk} - b_{21} - b_{22}x_{t2} - \ldots - b_{2m}x_{tm} + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{tk} = -a_{k1}y_{t1} - a_{k2}y_{t2} - \ldots - b_{k1} - b_{k2}x_{t2} - \ldots - b_{km}x_{tm} + \varepsilon_k. \end{cases} \tag{4.9}$$

Теперь хорошо видно, что переменная y_i одновременно является и зависимой (в i-м уравнении), и независимой (в остальных уравнениях).

Преобразуем исходную систему (4.8) к матричному виду. Введем для каждого момента времени t вектор-столбцы эндогенных переменных $\mathbf{y}_t$, предопределенных переменных $\mathbf{x}_t$, случайных ошибок $\boldsymbol{\varepsilon_t}$:

$$\mathbf{y}_t = \begin{pmatrix} y_{t1} \\ y_{t2} \\ \ldots \\ y_{tk} \end{pmatrix}, \ \mathbf{x}_t = \begin{pmatrix} 1 \\ x_{t2} \\ \ldots \\ x_{tm} \end{pmatrix}, \ \boldsymbol{\varepsilon_t} = \begin{pmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \ldots \\ \varepsilon_{tk} \end{pmatrix}.$$

Зададим также *постоянные матрицы*

$$\mathbf{A} = \begin{pmatrix} 1 & a_{12} & \ldots & a_{1k} \\ a_{21} & 1 & \ldots & a_{2k} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{k1} & a_{k2} & \ldots & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ b_{k1} & b_{k2} & \ldots & b_{km} \end{pmatrix}.$$

Система (4.8) записывается в виде n матричных уравнений:

$$\mathbf{A}\mathbf{y}_t + \mathbf{B}\mathbf{x}_t = \boldsymbol{\varepsilon_t} \quad (t \in \overline{1,n}). \tag{4.10}$$

Элементы матриц $\mathbf{A}, \mathbf{B}$ неизвестны и подлежат статистическому оцениванию. Их называют структурными коэффициентами модели. Будем предполагать, что матрица $\mathbf{A}$ неособая.

Выражая из структурной формы модели каждую эндогенную переменную через предопределенные, приходим к приведенной форме модели. В матричном виде приведенная форма запишется следующим образом:

$$\mathbf{y}_t = \mathbf{S}\mathbf{x}_t + \mathbf{u}_t. \tag{4.11}$$

Здесь матрица

$$\mathbf{S} = -\mathbf{A}^{-1}\,\mathbf{B} = (s_{ij}) \tag{4.12}$$

состоит из приведенных коэффициентов модели; $\mathbf{u}_t = \mathbf{A}^{-1}\boldsymbol{\varepsilon}_t$ – преобразованный вектор случайных ошибок.

Формула (4.12) выражает приведенные коэффициенты через структурные. Эта зависимость *нелинейна*, что обусловлено сложной нелинейной связью элементов матриц $\mathbf{A}^{-1}$ и $\mathbf{A}$. Число структурных коэффициентов, вообще говоря, *не совпадает* с числом приведенных коэффициентов.

4.4. Оценивание структурных коэффициентов систем одновременных уравнений

Каждое i-е уравнение приведенной системы является уравнением регрессии с зависимой переменной y_i:

$$y_{ti} = s_{i1} + s_{i2}x_{t2} + \ldots + s_{im}x_{tm} + u_{ti}. \tag{4.13}$$

При выполнении естественных условий на векторы $\boldsymbol{\varepsilon}_t$, аналогичных условиям Гаусса–Маркова, применение к уравнениям (4.13) обычного МНК дает несмещенные состоятельные оценки $\widehat{s}_{ij}$ приведенных коэффициентов. Тем самым определяется линейная функция регрессии каждой эндогенной переменной y_i на предопределенные переменные $x_1, x_2, \ldots, x_m$:

$$\widehat{y}_{ti} = \widehat{s}_{i1} + \widehat{s}_{i2}x_{t2} + \ldots + \widehat{s}_{im}x_{tm} \tag{4.14}$$

(мы рассматриваем эти функции лишь при выборочных значениях переменных).

В случае *идентифицируемости* исходной СОУ, когда можно *однозначно* выразить структурные коэффициенты через приведенные, применяют косвенный МНК. Для этого подставляют оценки $\widehat{s}_{ij}$ приведенных коэффициентов в формулы, определяющие структурные коэффициенты через приведенные. Тем самым находят несмещенные и состоятельные точечные оценки структурных коэффициентов.

Пусть теперь структурная модель *неидентифицируема в целом*. Тогда проблему идентифицируемости изучают применительно к *каждому уравнению* системы (4.8).

Уравнение структурной формы называется *идентифицируемым*, если *все* его коэффициенты *однозначно* определяются по приведенным коэффициентам.

Уравнение структурной формы называется *сверхидентифицируемым*, если все его коэффициенты определяются по приведенным коэффициентам, но хотя бы один из них выражается через приведенные коэффициенты *несколькими способами*.

Уравнение структурной формы называется *неидентифицируемым*, если хотя бы один его коэффициент *не может быть определен* через приведенные коэффициенты.

Если структурная форма содержит *неидентифицируемое* уравнение, то требуется *корректировка модели*.

Для оценивания *сверхидентифицируемого* уравнения применяют *двухшаговый метод наименьших квадратов* (2МНК). Опишем его.

Пусть i-е уравнение системы (4.8) сверхидентифицируемо. Перейдем к системе (4.9), тогда уравнение запишется в виде

$$\begin{aligned} y_{ti} = -a_{i1}y_{t1} - \ldots - a_{i,i-1}y_{t,i-1} - a_{i,i+1}y_{t,i+1} - \ldots - a_{ik}y_{tk} - \\ -b_{i1} - b_{i2}x_{t2} - \ldots - b_{im}x_{tm} + \varepsilon_i. \end{aligned} \tag{4.15}$$

На *первом* этапе 2МНК проводится оценивание всех регрессионных уравнений (4.13) и находятся функции регрессии (4.14).

На *втором* этапе метода, заменяя в правой части уравнения (4.15) эндогенные переменные их функциями регрессии, приходим к новому регрессионному уравнению

$$\begin{aligned} y_{ti} = -a_{i1}\widehat{y}_{t1} - \ldots - a_{i,i-1}\widehat{y}_{t,i-1} - a_{i,i+1}\widehat{y}_{t,i+1} - \ldots - a_{ik}\widehat{y}_{tk} - \\ -b_{i1} - b_{i2}x_{t2} - \ldots - b_{im}x_{tm} + \varepsilon_i. \end{aligned} \tag{4.16}$$

Применяя к уравнению (4.16) МНК, находим состоятельные оценки структурных коэффициентов уравнения (4.15).

Таким образом, в двухшаговом методе наименьших квадратов обычный МНК применяется дважды - на первом и втором этапах. При этом всякий раз оцениваются регрессионные уравнения, регрессорами в которых выступают предопределенные переменные.

Если оцениваемое уравнение *идентифицируемо*, то 2МНК приводит к тому же результату, что и косвенный МНК. Поэтому 2МНК – достаточно универсальный, широко применяемый на практике метод оценивания структурных коэффициентов СОУ.

Существует *трехшаговый метод наименьших квадратов* (3МНК), эффективный в случае коррелированности ошибок различных уравнений в СОУ. В нем два этапа 2МНК дополняются третьим, где с помощью ОМНК переоцениваются уже найденные оценки структурных коэффициентов.

Контрольные вопросы и задания

1. Обоснуйте необходимость рассмотрения систем одновременных уравнений.

2. Укажите различные типы переменных, входящих в системы одновременных уравнений.

3. Почему экзогенные и лаговые эндогенные переменные называют предопределенными?

4. Как взаимодействуют структурная и приведенная формы эконометрической модели?

5. В каких случаях и с какой целью применяется двухшаговый метод наименьших квадратов?

Заключение

В учебном пособии с позиций регрессионного анализа рассмотрены способы исследования основных типов эконометрических моделей. Ключевой задачей эконометрического исследования является статистическое оценивание коэффициентов модели. Для решения этой задачи применены метод наименьших квадратов и его обобщения. Автор надеется, что в результате усвоения базовых понятий эконометрики, изложенных в пособии, читатель сможет ориентироваться в проблематике эконометрического моделирования.

Конечно, данное пособие является лишь введением в обширный, разноплановый по методам исследования курс эконометрики.

Развернутая теория временных рядов опирается на аппарат случайных функций [1, гл. 16] .

Помимо метода наименьших квадратов в эконометрике применяются другие методы оценивания параметров: метод максимального правдоподобия [14, гл. 10], обобщенный метод моментов [14, с. 389–334]. Первый из них используют, например, при анализе регрессионных моделей с классификационными зависимыми переменными [14, гл. 12]. Обобщенный метод моментов находит применение при исследовании моделей с так называемыми панельными данными, сочетающими пространственные и временные данные [14, гл. 13].

Приложение

Некоторые элементы теории матриц

Матрицей размера $m \times n$ ($m \in \mathbb{N}$, $n \in \mathbb{N}$) называется прямоугольная таблица

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = (a_{ij}).$$

Числа a_{ij}, составляющие матрицу A, называются ее *элементами*; здесь

$$i \in \overline{1,m},\ j \in \overline{1,n}.$$

Матрица $\mathbf{A}$ имеет m строк и n столбцов. В обозначении a_{ij} элемента матрицы $\mathbf{A}$ i – номер строки, j – номер столбца, на пересечении которых стоит этот элемент.

Матрица $\mathbf{A}^T$ размера $n \times m$, получающаяся из матрицы $\mathbf{A}$ заменой каждой строки на столбец с тем же номером, называется матрицей, *транспонированной* к матрице $\mathbf{A}$. Укажем правила перехода к транспонированным матрицам:

$$(\mathbf{A}^T)^T = \mathbf{A},\ (\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T,\ (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \qquad (\text{П1})$$

Матрица размера $n \times n$ ($n \geqslant 2$) называется *квадратной матрицей порядка* n. Квадратная матрица $\mathbf{A} = (a_{ij})$ называется *симметрической*, если $\mathbf{A} = \mathbf{A}^T$, т.е. если $a_{ij} = a_{ji}$.

С помощью произвольной матрицы $\mathbf{A}$ размера $m \times n$ строятся симметрические матрицы: матрица $\mathbf{A}^T\mathbf{A}$ порядка n и матрица $\mathbf{A}\mathbf{A}^T$ порядка m. В частности, вектор-столбец размера $m \times 1$ $\mathbf{a} = (a_1, \dots, a_m)^T$ позволяет выделить две важные для приложений в эконометрике симметрические матрицы:

1) матрицу размера 1×1, т. е. число

$$\mathbf{a}^T\mathbf{a} = \sum_{i=1}^{m} a_i^2, \tag{П2}$$

являющееся суммой квадратов чисел a_i;

2) матрицу размера $m \times m$

$$\mathbf{a}\mathbf{a}^T = (a_i a_j) = \begin{pmatrix} a_1^2 & a_1a_2 & \dots & a_1a_m \\ a_1a_2 & a_2^2 & \dots & a_2a_m \\ \dots & \dots & \dots & \dots \\ a_1a_m & a_2a_m & \dots & a_m^2 \end{pmatrix}, \tag{П3}$$

составленную из всевозможных попарных произведений чисел a_i. На ее главной диагонали располагаются квадраты этих чисел.

Пусть теперь $\mathbf{A}$ — квадратная *неособая* матрица порядка n (т. е. ее определитель $\det \mathbf{A}$ не равен нулю). Она имеет обратную матрицу $\mathbf{A}^{-1}$, такую что

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{E}.$$

Здесь $\mathbf{E}$ — единичная матрица.

Справедлива формула

$$(\alpha\mathbf{A})^{-1} = \frac{1}{\alpha}\mathbf{A}^{-1} \quad (\alpha \neq 0). \tag{П4}$$

Обратная матрица вычисляется по формуле

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}((-1)^{i+j}M_{ij})^T. \tag{П5}$$

Здесь M_{ij} — минор элемента a_{ij} матрицы $\mathbf{A}$, т. е. определитель матрицы $(n-1)$-го порядка, получающейся из матрицы $\mathbf{A}$ вычеркиванием i-й строки и j-го столбца.

Для квадратной матрицы *второго* порядка

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

обратная матрица вычисляется проще:

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{П6}$$

Рассмотрим теперь *случайные матрицы*, элементы которых – случайные величины. Математическим ожиданием случайной матрицы $\mathbf{A} = \mathbf{A}_{m\times n} = (a_{ij})$ называют числовую матрицу размера $m \times n$, составленную из математических ожиданий $M(a_{ij})$ элементов матрицы $\mathbf{A}$:

$$M(\mathbf{A}) = (M(a_{ij})).$$

Свойства математического ожидания случайной матрицы аналогичны свойствам математического ожидания случайной величины. Например:

1) $M(\mathbf{A}_1 + \mathbf{A}_2) = M(\mathbf{A}_1) + M(\mathbf{A}_2)$; $\quad M(\alpha\mathbf{A}) = \alpha \cdot M(\mathbf{A})$ $(\alpha \in \mathbb{R})$;

2) если $\mathbf{C} = (c_{ij})$ постоянная матрица $(c_{ij} = const \in \mathbb{R})$, то $M(\mathbf{C}) = \mathbf{C}$, $M(\mathbf{CA}) = \mathbf{C} \cdot M(\mathbf{A})$.

Аналогом дисперсии для *случайного вектора*

$$\mathbf{a} = (a_1, \ldots, a_m)^T$$

служит числовая симметрическая матрица

$$\mathbf{K}(\mathbf{a}) = M((\mathbf{a} - M(\mathbf{a}))(\mathbf{a} - M(\mathbf{a}))^T).$$

Она состоит из элементов

$$M((a_i - M(a_i)) \cdot (a_j - M(a_j))) = cov(a_i, a_j),$$

являющихся ковариациями случайных величин a_i, a_j. Так как при $i = j$

$$cov(a_i, a_i) = D(a_i),$$

то на ее главной диагонали стоят дисперсии случайных величин a_i :

$$\mathbf{K}(\mathbf{a}) = \begin{pmatrix} D(a_1) & cov(a_1, a_2) & \ldots & cov(a_1, a_m) \\ cov(a_1, a_2) & D(a_2) & \ldots & cov(a_2, a_m) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ cov(a_1, a_m) & cov(a_2, a_m) & \ldots & D(a_m) \end{pmatrix}. \qquad (\Pi 7)$$

Матрицу $\mathbf{K}(\mathbf{a})$ называют *ковариационной матрицей случайного вектора* $\mathbf{a}$.

Рассмотрим случай, когда величины a_i, составляющие случайный вектор $\mathbf{a}$, неизвестны, а доступны лишь их выборочные значения

$$a_{1i},\ a_{2i},\ \ldots,\ a_{mi}.$$

Тогда вместо ковариаций $cov(a_i, a_j)$ между величинами a_i, a_j могут быть вычислены выборочные ковариации

$$k_{ij} = cov_{\text{в}}(a_i, a_j) = \overline{a_i a_j} - \overline{a}_i \cdot \overline{a}_j.$$

Здесь

$$\overline{a}_i = \frac{\sum\limits_{p=1}^{m} a_{pi}}{n}, \quad \overline{a_i a_j} = \frac{\sum\limits_{p=1}^{m} a_{pi} \cdot a_{pj}}{n}.$$

В частности, k_{ii} выборочная дисперсия величины a_i, причем $k_{ii} \geqslant 0$ ($k_{ii} = 0$ лишь в случае $a_{1i} = a_{2i} = \ldots = a_{mi}$). Поэтому вместо ковариационной матрицы $\mathbf{K}(\mathbf{a})$ можно ввести *выборочную ковариационную матрицу вектора* $\mathbf{a}$

$$\mathbf{K}_{\text{в}}(\mathbf{a}) = \begin{pmatrix} k_{11} & k_{12} & \ldots & k_{1m} \\ k_{12} & k_{22} & \ldots & k_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ k_{1m} & k_{2m} & \ldots & k_{mm} \end{pmatrix}. \qquad \text{(П8)}$$

Ее называют также выборочной ковариационной матрицей величин $a_1, a_2, \ldots, a_m$.

Нормируя элементы матрицы $\mathbf{K}_{\text{в}}(\mathbf{a})$, приходим к *(выборочной) корреляционной матрице случайного вектора* $\mathbf{a}$

$$\mathbf{R}(\mathbf{a}) = \mathbf{R} = (r_{ij}).$$

Ее элементы

$$r_{ij} = \frac{k_{ij}}{\sqrt{k_{ii} k_{jj}}}$$

являются выборочными коэффициентами корреляции между величинами a_i, a_j, причем $-1 \leqslant r_{ij} \leqslant 1$. Так как $r_{ii} = 1$, то подробнее корреляционная матрица записывается следующим образом:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{12} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1m} & r_{2m} & \ldots & 1 \end{pmatrix}. \qquad \text{(П9)}$$

Матрицы (П8), (П9) рассматривают и в случае, когда переменные $a_1, \ldots, a_m$, образующие вектор $\mathbf{a}$, или некоторые из них становятся

неслучайными, детерминированными. Если x_i, x_j – неслучайные переменные, то коэффициенты k_{ij}, r_{ij} являются характеристиками детерминированных связей между ними.

Градиент скалярной функции векторного аргумента

Пусть $\mathbf{x}$ – вектор-столбец переменных $x_1, x_2, \ldots, x_n$:

$$\mathbf{x} = (x_1,\ x_2,\ \ldots, x_n)^T.$$

Рассмотрим скалярную функцию векторного аргумента $Q(\mathbf{x})$. Иначе говоря, это – функция n переменных $Q(x_1, x_2, \ldots, x_n)$. Назовем важные простые классы таких функций:

1) *линейная функция* (без свободного члена)

$$\mathbf{a}^T\mathbf{x} = \mathbf{x}^T\mathbf{a} = \sum_{i=1}^{n} a_i x_i;$$

2) *квадратичная форма*

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i,j=1}^{n} a_{ij} x_i x_j.$$

Здесь $\mathbf{A}$ – квадратная симметрическая матрица.

Градиентом функции $Q(\mathbf{x})$ называют вектор-столбец, составленный из ее частных производных первого порядка:

$$\mathbf{grad}\ Q(\mathbf{x}) = \left(\frac{\partial Q(\mathbf{x})}{\partial x_1},\ \frac{\partial Q(\mathbf{x})}{\partial x_2},\ \ldots,\ \frac{\partial Q(\mathbf{x})}{\partial x_n}\right)^T. \qquad (\text{П10})$$

Укажем простейшие свойства градиента:

1. $\mathbf{grad}\ (cQ(\mathbf{x})) = c \cdot \mathbf{grad}\ Q(\mathbf{x})$ $\quad (c = const \in \mathbb{R})$; 2. $\mathbf{grad}\ (Q_1(\mathbf{x}) + Q_2(\mathbf{x})) = \mathbf{grad}\ Q_1(\mathbf{x}) + \mathbf{grad}\ Q_2(\mathbf{x})$.

Приведем формулы вычисления градиента линейной функции и квадратичной формы:

$$\mathbf{grad}\ \mathbf{a}^T\mathbf{x} = \mathbf{grad}\ \mathbf{x}^T\mathbf{a} = \mathbf{a}, \qquad (\text{П11})$$

$$\mathbf{grad}\ \mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}. \qquad (\text{П12})$$

Учебное издание

Густомесов Валерий Алексеевич

ЭКОНОМЕТРИКА

Учебное пособие

Редактор Н.М. Юркова
Компьютерная верстка В.А. Густомесова

Печатается по постановлению
редакционно-издательского совета
университета

Подписано в печать 13.12.04. Формат 60x84 1/16.
Бумага для множ. аппаратов. Печать плоская. Усл. п. л 7,4.
Тираж 200 экз. Заказ №241-Э

Отпечатано с готового оригинал-макета в типографии АМБ
620144, г. Екатеринбург, ул. Фрунзе, 96.
Тел.: 251-65-96, 251-66-04, 269-55-74